# Análise de regressão linear simples

Departamento de Matemática
Escola Superior de Tecnologia de Viseu

## Introdução

A análise de regressão estuda o relacionamento entre uma variável chamada a 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.

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.

# Introdução

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.

Portanto este diagrama permite decidir empiricamente se um relacionamento linear entre X e Y deve ser assumido.

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 recta imaginária que passa através do enxame de pontos.

# Diagrama de dispersão

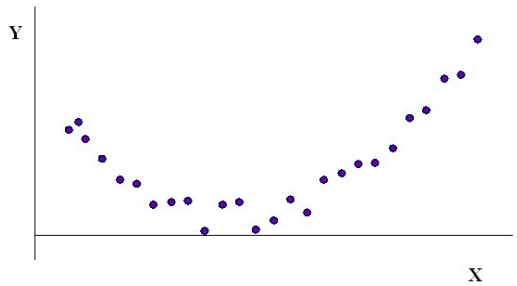
A correlação é tanto maior quanto mais os pontos se concentram, com pequenos desvios, em relação a essa recta.

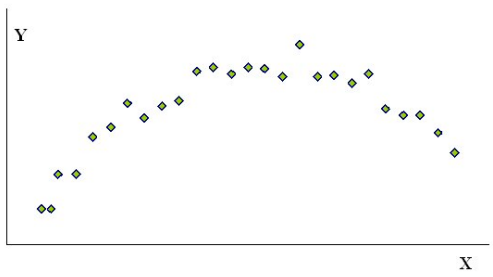
Se o declive da recta é positivo, concluímos que a correlação entre X e Y é positiva, i.e., os fenómenos variam no mesmo sentido.

Ao contrário, se o declive é negativo, então a correlação entre X e Y é negativa, i.e., os fenómenos variam em sentido inverso.

# Diagrama de dispersão

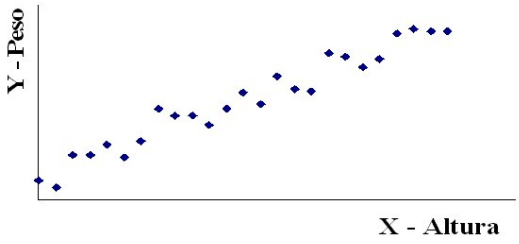
Sugerem uma regressão não linear
(i.e., a relação entre as duas variáveis poderá ser descrita por uma equação não linear)

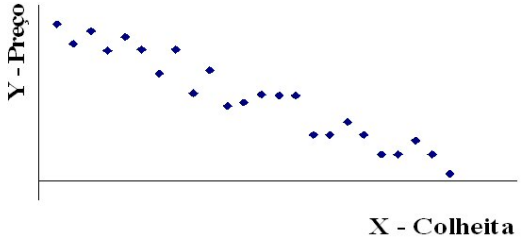


# Diagrama de dispersão

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



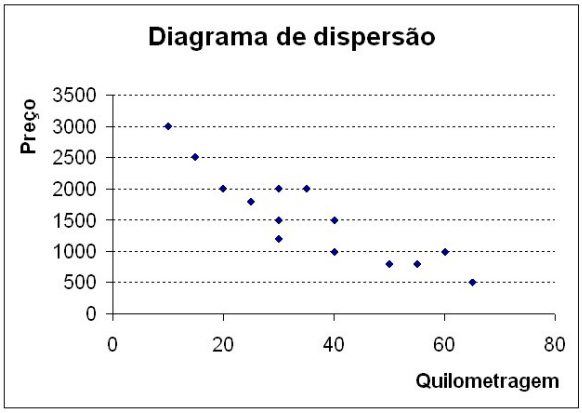

Existência de correlação positiva (em média, quanto maior for a altura maior será o peso)

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

# Exemplo

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.

# O 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;
$\beta_0$ e $\beta_1$ – parâmetros desconhecidos do modelo (a estimar);
Y – a 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)

---

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 $\sigma^2$.

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 $\sigma^2$

Os dados para a análise de regressão e correlação simples são da forma: $(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$ onde $x_i$ é o valor da variável $X$ e $y_i$ a correspondente observação da variável aleatória $Y_i$ $(i = 1, ..., n)$.

Cada observação satisfaz a seguinte relação:

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\mu_{Y_i}} + \varepsilon_i \qquad i = 1, ..., n$$

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

---

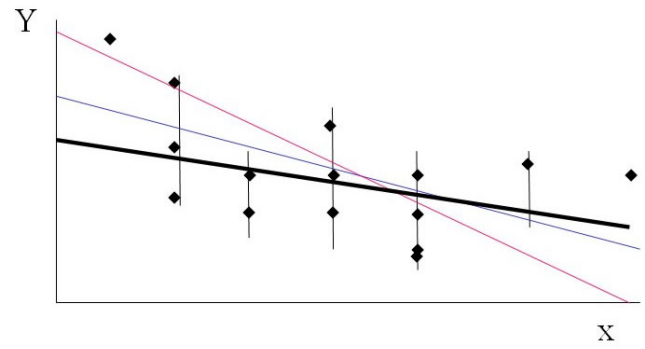
A partir dos dados disponíveis estimamos $\beta_0$ e $\beta_1$ e substituímos estes parâmetros pelas suas estimativas para obter a equação de regressão estimada.

$$\hat{y} = \hat{\mu}_{Y|x} = b_0 + b_1 x$$

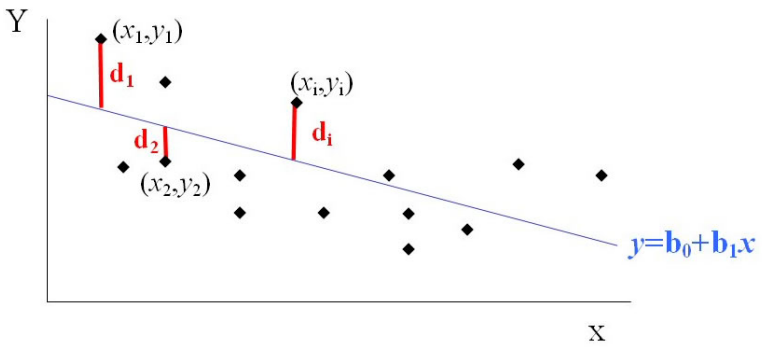
$\hookrightarrow$ Esta equação estima o valor médio de $Y$ para um dado valor $x$ de $X$, mas é usada para estimar o próprio valor de $Y$.

$\hookrightarrow$ 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 $\hat{\mu}_{Y|x}$.

# 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

$d_i=y_i-(b_0+b_1x_i)$

**resíduos**

# Estimação pelo método dos mínimos quadrados

Iremos estimar os parâmetros usando o método dos mínimos quadrados.

Seja $d_i = y_i - \hat{y}_i \hookrightarrow$ i-ésimo resíduo.

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

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

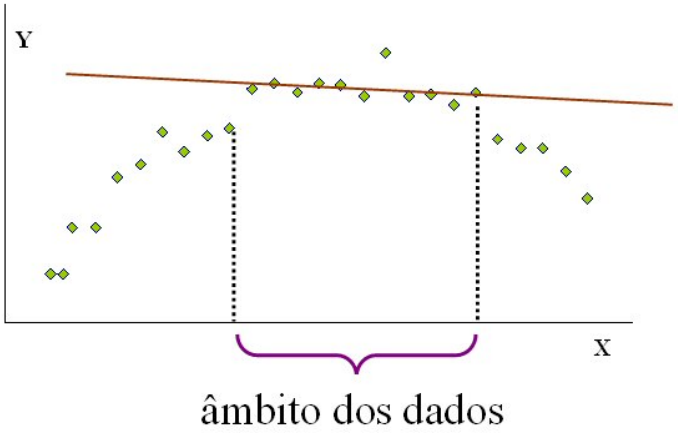
# Estimação pelo método dos mínimos quadrados

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.

# Exemplo - Estimação dos coeficientes de regressão

| 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** |

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

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

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^{\,2} - n\,\bar{x}^{\,2}} = \frac{640000 - 14 \times 36.07 \times 1542.85}{21825 - 14 \times 36.07^2} = -38.56$$

$$b_0 = \bar{y} - b_1\bar{x} = 1542.85 - 38.56 \times 36.07 = 2934$$

# Exemplo - Estimação dos coeficientes de regressão

Recta de regressão estimada: $\hat{y} = 2934 - 38.56x$

O preço esperado para um carro é de 2934 dezenas de Euros, menos 38.56 dezenas de Euros por cada mil Km que o carro tenha andado.

Por exemplo, para um carro que tenha andado 20000 Km, a equação sugere o preço:
$\hat{y} = 2934 - 38.5 \times 20 = 2162.8$ dezenas de Euros

O coeficiente de regressão estimado $b_1$ (estimativa de $\beta_1$), estima o efeito sobre o valor médio da variável dependente $Y$ de uma alteração unitária da variável independente $X$.

Assim, em média, por cada 1000 km que o carro tenha andado, o preço de venda baixa 38.56 dezenas de Euros.

# Exemplo - Estimação dos coeficientes de regressão

## Atenção

- $b_0 = 2934$ 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.
- Trata-se de uma relação média, assim um carro com determinada quilometragem não obterá necessariamente o preço de venda exacto indicado pela equação

# Qualidade do Ajustamento - Coeficiente de Correlação e de determinação

A equação de regressão estimada pode ser vista como uma tentativa para explicar as variações na variável dependente Y que resultam das alterações na variável dependente X.

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

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

# Qualidade do Ajustamento - Coeficiente de Correlação e de determinação

Isto é, se as predições baseadas na recta não são melhores que as baseadas no valor médio $\overline{y}$, então não adianta dispormos de uma equação de regressão.

Se a dispersão (erro) associada à recta é muito menor que a dispersão (erro) associada a $\overline{y}$, as predições baseadas na recta serão melhores que as baseadas em $\overline{y}$.

Dispersão em torno de $\overline{y}$ - Variação total:

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 \quad \text{(Soma dos quadrados totais)}$$

Dispersão em torno da recta de regressão - Variação não explicada:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{(Soma dos quadrados dos resíduos)}$$

O ajustamento será tanto melhor quanto mais pequeno for $SSE$ relativamente a $SST$.

Pode-se mostrar que:

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i-\overline{y})^2 & = & \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 & + & \sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 \\ \downarrow & & \downarrow & & \downarrow \\ SST & = & SSE & + & SSR \end{array}$$

$SST \longmapsto$ 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 |

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

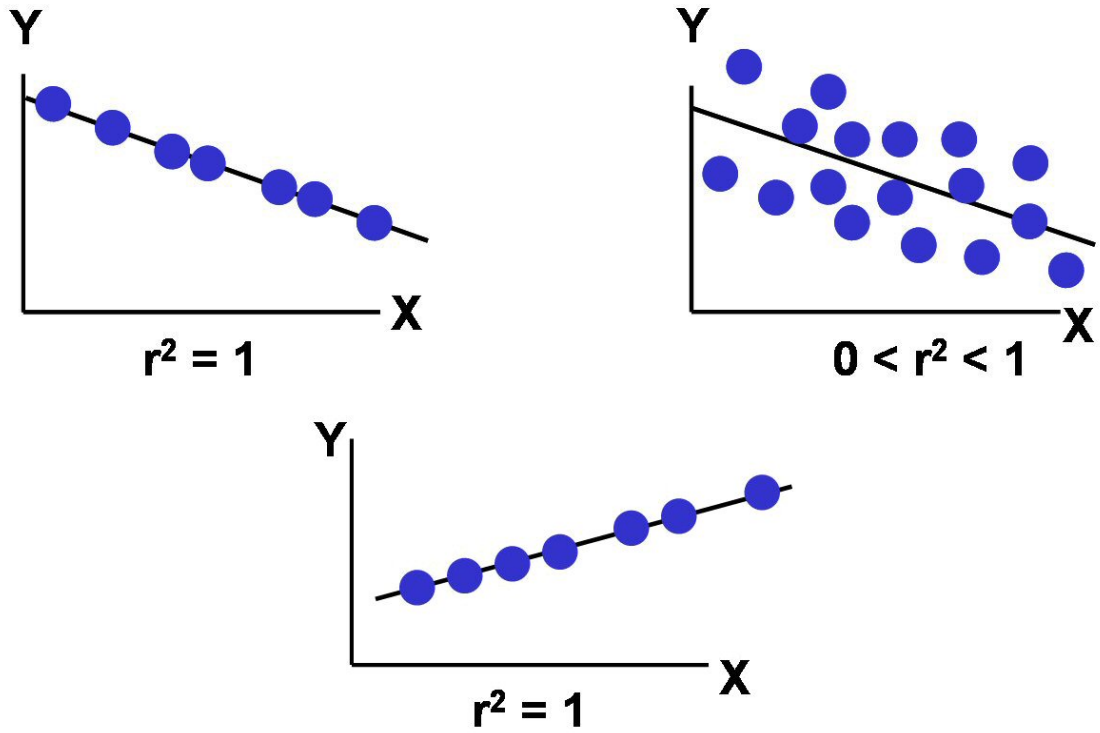
Este coeficiente pode ser utilizado como uma medida da qualidade do ajustamento, ou como medida da confiança depositada na equação de regressão como instrumento de previsão:

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

$r^2$ pode ser calculado a partir da seguinte fórmula:

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

## Exemplos de diagramas



# Coeficiente de Correlação e de determinação(3)

Á raiz quadrada de $r^2$ dá-se o nome de coeficiente de correlação simples.

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

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

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

# Exemplo - Determinação dos coeficientes de correlação e determinação

| 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** |

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

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

$b_0 = 2934$

$b_1 = -38.56$

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

$r^2 = 0.809 \longmapsto$
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 factores 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.

$r = -\sqrt{0.809} = -0.899 \longmapsto$
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.