Medidas de associação II

Correlação e Regressão Linear Simples

Felipe Figueiredo

Instituto Nacional de Traumatologia e Ortopedia

Sumário

Sumário

Tipos de variáveis envolvidas

- Considere duas amostras X e Y, de dados numéricos contínuos.
- Vamos representar os dados em pares ordenados (x,y) onde:
  - X: variável independente (ou variável explanatória)
  - Y: variável dependente (ou variável resposta)

Tipos de variáveis envolvidas

- Considere duas amostras X e Y, de dados numéricos contínuos.
- Vamos representar os dados em pares ordenados (x,y) onde:
  - X: variável independente (ou variável explanatória)
  - Y: variável dependente (ou variável resposta)

Tipos de variáveis envolvidas

- Considere duas amostras X e Y, de dados numéricos contínuos.
- Vamos representar os dados em pares ordenados (x,y) onde:
    - X: variável independente (ou variável explanatória)
    - Y: variável dependente (ou variável resposta)

Tipos de variáveis envolvidas

- Considere duas amostras X e Y, de dados numéricos contínuos.
- Vamos representar os dados em pares ordenados (x,y) onde:
  - X: variável independente (ou variável explanatória)
  - Y: variável dependente (ou variável resposta)

Medidas de associação

- Como definir (e mensurar!) o grau de associação entre duas variáveis aleatórias (VAs)?
- Se uma VA é dependente de outra, é razoável assumir que isso possa ser observável por estatísticas sumárias
- Como resumir esta informação em uma única grandeza numérica?

Medidas de associação

- Como definir (e mensurar!) o grau de associação entre duas variáveis aleatórias (VAs)?
- Se uma VA é dependente de outra, é razoável assumir que isso possa ser observável por estatísticas sumárias
- Como resumir esta informação em uma única grandeza numérica?

Medidas de associação

- Como definir (e mensurar!) o grau de associação entre duas variáveis aleatórias (VAs)?
- Se uma VA é dependente de outra, é razoável assumir que isso possa ser observável por estatísticas sumárias
- Como resumir esta informação em uma única grandeza numérica?

Medidas de associação

- Quando uma associação é forte, podemos identificá-la subjetivamente
- Para isto, analisamos o gráfico de dispersão dos pares (x,y)
- Um gráfico deste tipo é feito simplesmente plotando os pontos no plano cartesiano

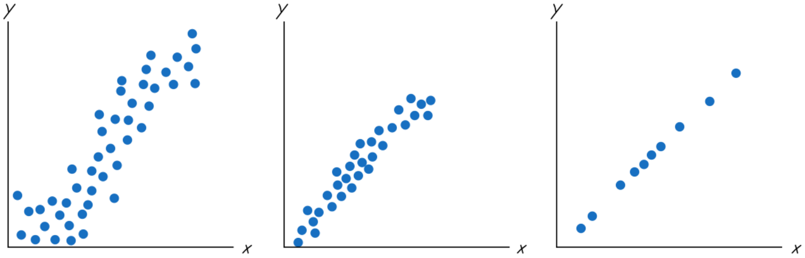
Medidas de associação

- Quando uma associação é forte, podemos identificá-la subjetivamente
- Para isto, analisamos o gráfico de dispersão dos pares (x,y)
- Um gráfico deste tipo é feito simplesmente plotando os pontos no plano cartesiano

Medidas de associação

- Quando uma associação é forte, podemos identificá-la subjetivamente
- Para isto, analisamos o gráfico de dispersão dos pares (x,y)
- Um gráfico deste tipo é feito simplesmente plotando os pontos no plano cartesiano

Exemplo



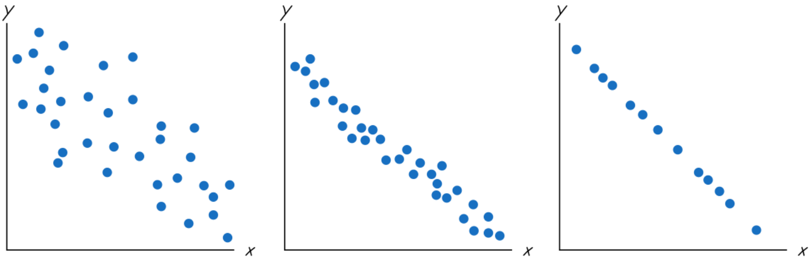
(a) Positive correlation between x and y

(b) Strong positive correlation between x and y

(c) Perfect positive correlation between x and y

(Fonte: Triola)

Exemplo



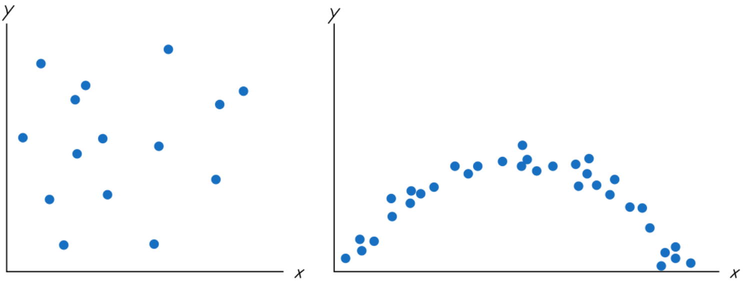
(d) Negative correlation between x and y

(e) Strong negative correlation between x and y

(f) Perfect negative correlation between x and y

(Fonte: Triola)

Exemplo



(g) No correlation between x and y

(h) Nonlinear relationship between x and y

(Fonte: Triola)

Sumário

Variância

- Relembrando: a variância (assim como o desvio-padrão) é uma medida da dispersão da amostra
- Medida sumária que resume o quanto os dados se desviam da média
- Podemos usar um raciocínio análogo para comparar quanto uma amostra se desvia em relação à outra

Variância

- Relembrando: a variância (assim como o desvio-padrão) é uma medida da dispersão da amostra
- Medida sumária que resume o quanto os dados se desviam da média
- Podemos usar um raciocínio análogo para comparar quanto uma amostra se desvia em relação à outra

Variância

- Relembrando: a variância (assim como o desvio-padrão) é uma medida da dispersão da amostra
- Medida sumária que resume o quanto os dados se desviam da média
- Podemos usar um raciocínio análogo para comparar quanto uma amostra se desvia em relação à outra

Covariância entre duas amostras

Definition

A covariância entre duas variáveis X e Y é uma medida de quanto ambas variam juntas (uma em relação à outra).

- Obs: duas variáveis independentes tem covariância igual a zero!

Covariância entre duas amostras

Definition

A covariância entre duas variáveis X e Y é uma medida de quanto ambas variam juntas (uma em relação à outra).

- Obs: duas variáveis independentes tem covariância igual a zero!

Correlação

Definition

A correlação é a associação estatística entre duas variáveis.

Para medir essa associação, calculamos o coeficiente de correlação r.

Sumário

Coeficiente de correlação

Definition

O coeficiente de correlação r é a medida da direção e força da associação entre duas variáveis.

Propriedades:

- É um número entre -1 e 1.
- Mede a associação linear entre duas variáveis.
  - Diretamente proporcional, inversamente proporcional, ou ausência de proporcionalidade.

Coeficiente de correlação

Definition

O coeficiente de correlação r é a medida da direção e força da associação entre duas variáveis.

Propriedades:

- É um número entre -1 e 1.
- Mede a associação linear entre duas variáveis.
  - Diretamente proporcional, inversamente proporcional, ou ausência de proporcionalidade.

Coeficiente de correlação

Definition

O coeficiente de correlação r é a medida da direção e força da associação entre duas variáveis.

Propriedades:

- É um número entre -1 e 1.
- Mede a associação linear entre duas variáveis.
  - Diretamente proporcional, inversamente proporcional, ou ausência de proporcionalidade.

Coeficiente de correlação

Definition

O coeficiente de correlação r é a medida da direção e força da associação entre duas variáveis.

Propriedades:

- É um número entre -1 e 1.
- Mede a associação linear entre duas variáveis.
    - Diretamente proporcional, inversamente proporcional, ou ausência de proporcionalidade.

Coeficiente de correlação

- O coeficiente de correlação de Pearson é a covariância normalizada
- Pode ser calculado para populações (ρ) ou amostras (r)
- População

$$\rho = \frac{\text{Cov(X,Y)}}{\sigma_X \times \sigma_Y}$$

- Utilizando uma fórmula semelhante, encontramos o coeficiente r para uma amostra

Coeficiente de correlação

- O coeficiente de correlação de Pearson é a covariância normalizada
- Pode ser calculado para populações (ρ) ou amostras (r)
- População

$$\rho = \frac{\text{Cov(X,Y)}}{\sigma_X \times \sigma_Y}$$

- Utilizando uma fórmula semelhante, encontramos o coeficiente r para uma amostra

Coeficiente de correlação

- O coeficiente de correlação de Pearson é a covariância normalizada
- Pode ser calculado para populações (ρ) ou amostras (r)
- População

$$\rho = \frac{\text{Cov(X,Y)}}{\sigma_X \times \sigma_Y}$$

- Utilizando uma fórmula semelhante, encontramos o coeficiente r para uma amostra

Coeficiente de correlação

- O coeficiente de correlação de Pearson é a covariância normalizada
- Pode ser calculado para populações (ρ) ou amostras (r)
- População

$$\rho = \frac{\text{Cov(X,Y)}}{\sigma_X \times \sigma_Y}$$

- Utilizando uma fórmula semelhante, encontramos o coeficiente r para uma amostra

Correlação

- Uma forte associação positiva corresponde a uma correlação próxima de 1.
- Uma forte associação negativa corresponde a uma correlação próxima de -1.
- A ausência de associação corresponde a uma correlação próxima de 0.

Correlação

- Uma forte associação positiva corresponde a uma correlação próxima de 1.
- Uma forte associação negativa corresponde a uma correlação próxima de -1.
- A ausência de associação corresponde a uma correlação próxima de 0.

Correlação

- Uma forte associação positiva corresponde a uma correlação próxima de 1.
- Uma forte associação negativa corresponde a uma correlação próxima de -1.
- A ausência de associação corresponde a uma correlação próxima de 0.

Correlação

- Uma forte associação positiva corresponde a uma correlação próxima de 1.
- Uma forte associação negativa corresponde a uma correlação próxima de -1.
- A ausência de associação corresponde a uma correlação próxima de 0.

- Se tivéssemos os dados de toda a população, poderíamos calcular o parâmetro ρ
- Na prática, só podemos calcular a estatística r da amostra
- Utilizamos r como estimador para ρ, e testamos a significância estatística da forma usual

- Se tivéssemos os dados de toda a população, poderíamos calcular o parâmetro ρ
- Na prática, só podemos calcular a estatística r da amostra
- Utilizamos r como estimador para ρ, e testamos a significância estatística da forma usual

- Se tivéssemos os dados de toda a população, poderíamos calcular o parâmetro ρ
- Na prática, só podemos calcular a estatística r da amostra
- Utilizamos r como estimador para ρ, e testamos a significância estatística da forma usual

Exemplo

Example

Pesquisadores queriam entender por que a insulina varia tanto entre indivíduos. Imaginaram que a composição lipídica das células do músculo afetam a sensibilidade do músculo para a insulina. Para isto, eles injetaram insulina em 13 jovens adultos, e determinaram quanta glicose eles precisariam injetar nos sujeitos para manter o nível de glicose sanguínea constante. A quantidade de glicose injetada para manter o nível sanguíneo constante é, então, uma medida da sensibilidade à insulina.
(Fonte: Motulsky, 1995)

Exemplo

Example

Os pesquisadores fizeram uma pequena biópsia nos músculos para aferir a fração de ácidos graxos poliinsaturados que tem entre 20 e 22 carbonos (%C20-22). Como variável resposta, mediram o índice de sensibilidade à insulina.

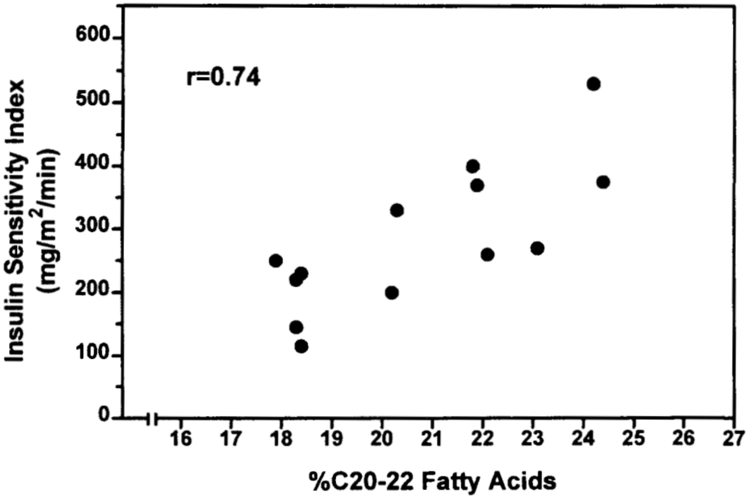
Valores tabelados a seguir.

Exemplo

Table 17.1. Correlation Between %C20–22 and Insulin Sensitivity

% C20–22 Polyunsaturated Fatty Acids	Insulin Sensitivity (mg/m^2/min)
17.9	250
18.3	220
18.3	145
18.4	115
18.4	230
20.2	200
20.3	330
21.8	400
21.9	370
22.1	260
23.1	270
24.2	530
24.4	375

Exemplo: Diagrama de dispersão dos dados



Obs: na verdade, $r = 0.77$.

Exemplo

- O tamanho da amostra foi $n = 13$
- Consultamos o valor crítico de r na tabela a seguir
- Testamos a H_0 que não há relação entre as variáveis na população ($H_0 : \rho = 0$).

Exemplo

- O tamanho da amostra foi $n = 13$
- Consultamos o valor crítico de r na tabela a seguir
- Testamos a H_0 que não há relação entre as variáveis na população ($H_0 : \rho = 0$).

Exemplo

- O tamanho da amostra foi $n = 13$
- Consultamos o valor crítico de r na tabela a seguir
- Testamos a H_0 que não há relação entre as variáveis na população ($H_0 : \rho = 0$).

Exemplo

TABLE A-6 Critical Values of the Pearson Correlation Coefficient r

n	$\alpha = .05$	$\alpha = .01$
4	.950	.999
5	.878	.959
6	.811	.917
7	.754	.875
8	.707	.834
9	.666	.798
10	.632	.765
11	.602	.735
12	.576	.708
13	.553	.684
14	.532	.661
15	.514	.641
16	.497	.623
17	.482	.606
18	.468	.590

Exemplo

- O valor crítico da tabela para uma amostra de tamanho 13 é $r_c = 0.553$
- A correlação calculada para esta amostra foi $r = 0.77$
- Como a correlação é maior que o valor crítico, a relação é estatisticamente significativa
- Conclusão: há evidências para rejeitar a H_0 que não há relação entre as variáveis.

Exemplo

- O valor crítico da tabela para uma amostra de tamanho 13 é $r_c = 0.553$
- A correlação calculada para esta amostra foi $r = 0.77$
- Como a correlação é maior que o valor crítico, a relação é estatisticamente significativa
- Conclusão: há evidências para rejeitar a H_0 que não há relação entre as variáveis.

Exemplo

- O valor crítico da tabela para uma amostra de tamanho 13 é $r_c = 0.553$
- A correlação calculada para esta amostra foi $r = 0.77$
- Como a correlação é maior que o valor crítico, a relação é estatisticamente significativa
- Conclusão: há evidências para rejeitar a H_0 que não há relação entre as variáveis.

Exemplo

- O valor crítico da tabela para uma amostra de tamanho 13 é $r_c = 0.553$
- A correlação calculada para esta amostra foi $r = 0.77$
- Como a correlação é maior que o valor crítico, a relação é estatisticamente significativa
- Conclusão: há evidências para rejeitar a H_0 que não há relação entre as variáveis.

Exemplo

- Pode-se também calcular o p-valor para o coeficiente de correlação *r*.
- Para este exemplo, teríamos $p = 0.0021$.
- Interpretação: se não houver relação entre as variáveis (H_0), existe apenas 0.21% de chance de observamos uma correlação tão forte com um estudo deste tamanho

Exemplo

- Pode-se também calcular o p-valor para o coeficiente de correlação r.
- Para este exemplo, teríamos $p = 0.0021$.
- Interpretação: se não houver relação entre as variáveis (H_0), existe apenas 0.21% de chance de observamos uma correlação tão forte com um estudo deste tamanho

Exemplo

- Pode-se também calcular o p-valor para o coeficiente de correlação r.
- Para este exemplo, teríamos $p = 0.0021$.
- Interpretação: se não houver relação entre as variáveis (H_0), existe apenas 0.21% de chance de observamos uma correlação tão forte com um estudo deste tamanho

Exemplo

Por que as duas variáveis são tão correlacionadas?
Considere 4 possibilidades:

1. o conteúdo lipídico das membranas determina a sensibilidade à insulina
2. A sensibilidade à insulina de alguma forma afeta o conteúdo lipídico
3. tanto o conteúdo lipídico quanto a sensibilidade à insulina estão sob o efeito de algum outro fator (talvez algum hormônio)
4. as duas variáveis não são correlacionads na população, e a estimativa observada nessa amostra é mera coincidência

Exemplo

Por que as duas variáveis são tão correlacionadas?
Considere 4 possibilidades:

1. o conteúdo lipídico das membranas determina a sensibilidade à insulina
2. A sensibilidade à insulina de alguma forma afeta o conteúdo lipídico
3. tanto o conteúdo lipídico quanto a sensibilidade à insulina estão sob o efeito de algum outro fator (talvez algum hormônio)
4. as duas variáveis não são correlacionads na população, e a estimativa observada nessa amostra é mera coincidência

Exemplo

Por que as duas variáveis são tão correlacionadas?
Considere 4 possibilidades:

1. o conteúdo lipídico das membranas determina a sensibilidade à insulina
2. A sensibilidade à insulina de alguma forma afeta o conteúdo lipídico
3. tanto o conteúdo lipídico quanto a sensibilidade à insulina estão sob o efeito de algum outro fator (talvez algum hormônio)
4. as duas variáveis não são correlacionads na população, e a estimativa observada nessa amostra é mera coincidência

Exemplo

Por que as duas variáveis são tão correlacionadas?
Considere 4 possibilidades:

1. o conteúdo lipídico das membranas determina a sensibilidade à insulina
2. A sensibilidade à insulina de alguma forma afeta o conteúdo lipídico
3. tanto o conteúdo lipídico quanto a sensibilidade à insulina estão sob o efeito de algum outro fator (talvez algum hormônio)
4. as duas variáveis não são correlacionads na população, e a estimativa observada nessa amostra é mera coincidência

Interpretando o r

- Nunca devemos ignorar a última possibilidade (erro tipo I)!
- o p-valor indica quão rara é essa coincidência
- neste caso, em apenas 0.21% dos experimentos não haveria uma correlação real, e estaríamos cometendo um erro de interpretação

Interpretando o r

- Nunca devemos ignorar a última possibilidade (erro tipo I)!
- o p-valor indica quão rara é essa coincidência
- neste caso, em apenas 0.21% dos experimentos não haveria uma correlação real, e estaríamos cometendo um erro de interpretação

Interpretando o r

- Nunca devemos ignorar a última possibilidade (erro tipo I)!
- o p-valor indica quão rara é essa coincidência
- neste caso, em apenas 0.21% dos experimentos não haveria uma correlação real, e estaríamos cometendo um erro de interpretação

Elevando o r ao quadrado

- Relembrando: calculamos a variância de uma amostra para saber a dispersão dos dados
- Sua interpretação é confusa, portanto preferimos usar o desvio-padrão
- No caso do r é o contrário: a interpretação de r^2 é mais simples
- Obs: o valor r^2 também é chamado coeficiente de determinação, como veremos a seguir.

Elevando o r ao quadrado

- Relembrando: calculamos a variância de uma amostra para saber a dispersão dos dados
- Sua interpretação é confusa, portanto preferimos usar o desvio-padrão
- No caso do r é o contrário: a interpretação de r^2 é mais simples
- Obs: o valor r^2 também é chamado coeficiente de determinação, como veremos a seguir.

Elevando o r ao quadrado

- Relembrando: calculamos a variância de uma amostra para saber a dispersão dos dados
- Sua interpretação é confusa, portanto preferimos usar o desvio-padrão
- No caso do r é o contrário: a interpretação de r^2 é mais simples
- Obs: o valor r^2 também é chamado coeficiente de determinação, como veremos a seguir.

Elevando o r ao quadrado

- Relembrando: calculamos a variância de uma amostra para saber a dispersão dos dados
- Sua interpretação é confusa, portanto preferimos usar o desvio-padrão
- No caso do r é o contrário: a interpretação de r^2 é mais simples
- Obs: o valor r^2 também é chamado coeficiente de determinação, como veremos a seguir.

Interpretando o r^2

- No exemplo anterior, $r^2 = 0.59$
- no caso, 59% da variabilidade da tolerância à insulina pode ser explicada pelo conteúdo lipídico
- Ou seja: conhecer o conteúdo lipídico permite explicar 59% da variância na sensibilidade à insulina
- Isto deixa 41% da variância que pode ser explicada por outros fatores ou erros de medição
- E este valor (r^2) também é utilizado na Regressão!

Interpretando o r^2

- No exemplo anterior, $r^2 = 0.59$
- no caso, 59% da variabilidade da tolerância à insulina pode ser explicada pelo conteúdo lipídico
- Ou seja: conhecer o conteúdo lipídico permite explicar 59% da variância na sensibilidade à insulina
- Isto deixa 41% da variância que pode ser explicada por outros fatores ou erros de medição
- E este valor (r^2) também é utilizado na Regressão!

Interpretando o r^2

- No exemplo anterior, $r^2 = 0.59$
- no caso, 59% da variabilidade da tolerância à insulina pode ser explicada pelo conteúdo lipídico
- Ou seja: conhecer o conteúdo lipídico permite explicar 59% da variância na sensibilidade à insulina
- Isto deixa 41% da variância que pode ser explicada por outros fatores ou erros de medição
- E este valor (r^2) também é utilizado na Regressão!

Interpretando o r^2

- No exemplo anterior, $r^2 = 0.59$
- no caso, 59% da variabilidade da tolerância à insulina pode ser explicada pelo conteúdo lipídico
- Ou seja: conhecer o conteúdo lipídico permite explicar 59% da variância na sensibilidade à insulina
- Isto deixa 41% da variância que pode ser explicada por outros fatores ou erros de medição
- E este valor (r^2) também é utilizado na Regressão!

Interpretando o r^2

- No exemplo anterior, $r^2 = 0.59$
- no caso, 59% da variabilidade da tolerância à insulina pode ser explicada pelo conteúdo lipídico
- Ou seja: conhecer o conteúdo lipídico permite explicar 59% da variância na sensibilidade à insulina
- Isto deixa 41% da variância que pode ser explicada por outros fatores ou erros de medição
- E este valor (r^2) também é utilizado na Regressão!

Sumário

1. Correlação
   - Associação entre duas variáveis
   - Covariância entre duas amostras
   - Coeficiente de correlação de Pearson
2. Regressão Linear Simples
   - Modelos estatísticos
   - Coeficiente de Determinação r^2
3. Interpretação
4. Causalidade
5. Resumo

Modelos estatísticos

Modelos servem para:

- representar de forma simplificada fenômenos, experimentos, dados, etc;
- possibilitar análise em cenários controlados, menos complexos que a realidade;
- extrapolar resultados e conclusões.

Modelos estatísticos

Modelos servem para:

- representar de forma simplificada fenômenos, experimentos, dados, etc;
- possibilitar análise em cenários controlados, menos complexos que a realidade;
- extrapolar resultados e conclusões.

Modelos estatísticos

Modelos servem para:

- representar de forma simplificada fenômenos, experimentos, dados, etc;
- possibilitar análise em cenários controlados, menos complexos que a realidade;
- extrapolar resultados e conclusões.

Modelos estatísticos

Ao ajustar um modelo aos dados, podemos:

- fazer predições dentro do intervalo observado para dados que não foram obtidos (interpolação)
- fazer predições fora do intervalo observado (extrapolação)

Modelos estatísticos

Ao ajustar um modelo aos dados, podemos:

- fazer predições dentro do intervalo observado para dados que não foram obtidos (interpolação)
- fazer predições fora do intervalo observado (extrapolação)

Reta de regressão

Definition

Uma reta de regressão (também chamada de reta de melhor ajuste) é a reta para a qual a soma dos erros quadráticos dos resíduos é o mínimo.

- É a reta que melhor se ajusta aos dados
- Minimiza os resíduos

Reta de regressão

Definition

Uma reta de regressão (também chamada de reta de melhor ajuste) é a reta para a qual a soma dos erros quadráticos dos resíduos é o mínimo.

- É a reta que melhor se ajusta aos dados
- Minimiza os resíduos

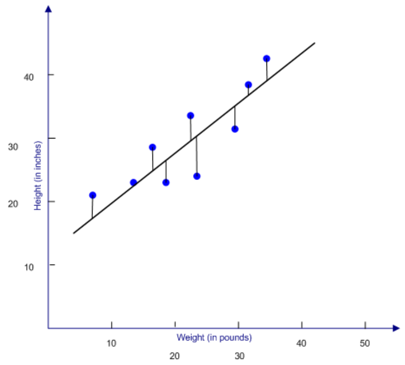
Reta de regressão

Definition

Uma reta de regressão (também chamada de reta de melhor ajuste) é a reta para a qual a soma dos erros quadráticos dos resíduos é o mínimo.

- É a reta que melhor se ajusta aos dados
- Minimiza os resíduos

Resíduos



Definition

Resíduos são a distância entre o dado observado e a reta estimada (modelo).

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula

$$\hat{y} = ax + b$$

- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula
$$\hat{y} = ax + b$$
- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula
$$\hat{y} = ax + b$$
- No caso da reta regressora:
    - y é a variável dependente
    - x é a variável independente
    - a é a inclinação
    - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula
$$\hat{y} = ax + b$$
- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula
$$\hat{y} = ax + b$$
- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula
$$\hat{y} = ax + b$$
- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Elementos da reta de regressão

- Relembrando: a equação de uma reta é definida pela fórmula
$$\hat{y} = ax + b$$
- No caso da reta regressora:
  - y é a variável dependente
  - x é a variável independente
  - a é a inclinação
  - b é o intercepto
- Assim, o objetivo da análise de regressão é encontrar os valores a e b

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- . . . e algumas operações entre estes termos

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- ...e algumas operações entre estes termos

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- ...e algumas operações entre estes termos

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

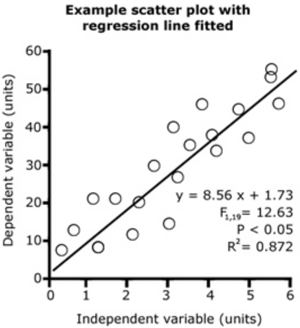
- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- ...e algumas operações entre estes termos

Análise de Regressão

Para determinar a inclinação e o intercepto, usamos:

- as médias de X e Y
- as variâncias de X e Y
- o coeficiente de correlação r entre X e Y
- o tamanho da amostra n
- ...e algumas operações entre estes termos

Análise de Regressão



- A qualidade do ajuste do modelo de regressão é determinado pelo coeficiente de determinação r^2

Sumário

Coeficiente de Determinação r^2

Definition

O coeficiente de determinação r^2 é a relação da variação explicada com a variação total.

$$r^2 = \frac{\text{variação explicada}}{\text{variação total}}$$

- Lembrando: r^2 é o quadrado de r!

Coeficiente de Determinação r^2

Definition

O coeficiente de determinação r^2 é a relação da variação explicada com a variação total.

$$r^2 = \frac{\text{variação explicada}}{\text{variação total}}$$

- Lembrando: r^2 é o quadrado de r!

Coeficiente de Determinação r^2

- Qual é a porcentagem da variação dos dados pode ser explicada pela reta regressora?
- O coeficiente r^2 é a fração da variância que é compartilhada entre X e Y.
- Como r está sempre entre -1 e 1, r^2 está sempre entre 0 e 1.

Coeficiente de Determinação r^2

- Qual é a porcentagem da variação dos dados pode ser explicada pela reta regressora?
- O coeficiente r^2 é a fração da variância que é compartilhada entre X e Y.
- Como r está sempre entre -1 e 1, r^2 está sempre entre 0 e 1.

Coeficiente de Determinação r^2

- Qual é a porcentagem da variação dos dados pode ser explicada pela reta regressora?
- O coeficiente r^2 é a fração da variância que é compartilhada entre X e Y.
- Como r está sempre entre -1 e 1, r^2 está sempre entre 0 e 1.

Coeficiente de Determinação r^2

- Além disso, $r^2 \leq |r|$
- Por que?

Compare os seguintes números entre 0 e 1:

$$\frac{1}{2} \text{ e } \left(\frac{1}{2}\right)^2 = \frac{1}{4} \Rightarrow \frac{1}{4} \leq \frac{1}{2}$$

$$\frac{1}{3} \text{ e } \left(\frac{1}{3}\right)^2 = \frac{1}{9} \Rightarrow \frac{1}{9} \leq \frac{1}{3}$$

Coeficiente de Determinação r^2

- Além disso, $r^2 \leq |r|$
- Por que?

Compare os seguintes números entre 0 e 1:

$$\frac{1}{2} \text{ e } \left(\frac{1}{2}\right)^2 = \frac{1}{4} \Rightarrow \frac{1}{4} \leq \frac{1}{2}$$

$$\frac{1}{3} \text{ e } \left(\frac{1}{3}\right)^2 = \frac{1}{9} \Rightarrow \frac{1}{9} \leq \frac{1}{3}$$

Coeficiente de Determinação r^2

- Além disso, $r^2 \leq |r|$
- Por que?

Compare os seguintes números entre 0 e 1:

$$\frac{1}{2} \text{ e } \left(\frac{1}{2}\right)^2 = \frac{1}{4} \Rightarrow \frac{1}{4} \leq \frac{1}{2}$$

$$\frac{1}{3} \text{ e } \left(\frac{1}{3}\right)^2 = \frac{1}{9} \Rightarrow \frac{1}{9} \leq \frac{1}{3}$$

Interpretação

- Se a correlação é 0, então X e Y não variam juntos (independentes)
- Se a correlação é positiva, então quando uma aumenta, a outra aumenta em proporção direta (linear)
- Se a correlação é negativa, então quando uma aumenta, a outra diminui em proporção inversa (linear)

Interpretação

- Se a correlação é 0, então X e Y não variam juntos (independentes)
- Se a correlação é positiva, então quando uma aumenta, a outra aumenta em proporção direta (linear)
- Se a correlação é negativa, então quando uma aumenta, a outra diminui em proporção inversa (linear)

Interpretação

- Se a correlação é 0, então X e Y não variam juntos (independentes)
- Se a correlação é positiva, então quando uma aumenta, a outra aumenta em proporção direta (linear)
- Se a correlação é negativa, então quando uma aumenta, a outra diminui em proporção inversa (linear)

Cuidado!

- Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável
- Ex: em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita
- Mas comprar mais telefones não vai salvar crianças!
- Explicação alternativa: a melhoria da condições financeiras pode afetar ambas as variáveis

Cuidado!

- Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável
- Ex: em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita
- Mas comprar mais telefones não vai salvar crianças!
- Explicação alternativa: a melhoria da condições financeiras pode afetar ambas as variáveis

Cuidado!

- Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável
- Ex: em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita
- Mas comprar mais telefones não vai salvar crianças!
- Explicação alternativa: a melhoria da condições financeiras pode afetar ambas as variáveis

Cuidado!

- Duas variáveis podem parecer correlacionadas pois são influenciadas por uma terceira variável
- Ex: em alguns países a mortalidade infantil é negativamente correlacionada com o número de telefones per capita
- Mas comprar mais telefones não vai salvar crianças!
- Explicação alternativa: a melhoria da condições financeiras pode afetar ambas as variáveis

Causa x efeito

- Se há uma relação de causalidade entre as duas variáveis, a correlação será não nula (positiva ou negativa)
- Quanto maior for a relação de dependência entre as variáveis, maior será o módulo da correlação.
- Se as variáveis não são relacionadas, a correlação será nula.

Causa x efeito

- Se há uma relação de causalidade entre as duas variáveis, a correlação será não nula (positiva ou negativa)
- Quanto maior for a relação de dependência entre as variáveis, maior será o módulo da correlação.
- Se as variáveis não são relacionadas, a correlação será nula.

Causa x efeito

- Se há uma relação de causalidade entre as duas variáveis, a correlação será não nula (positiva ou negativa)
- Quanto maior for a relação de dependência entre as variáveis, maior será o módulo da correlação.
- Se as variáveis não são relacionadas, a correlação será nula.

Causalidade?

- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Causalidade?

- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Causalidade?

- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Causalidade?

- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Causalidade?

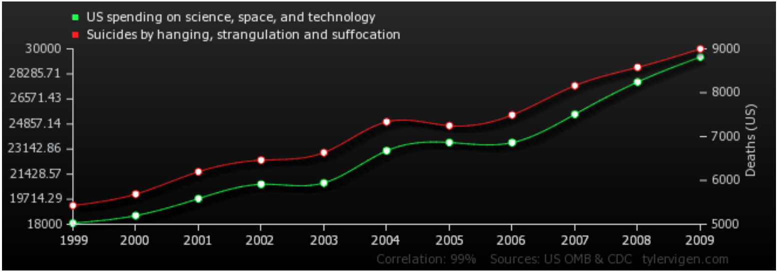
- Mas não podemos inverter a afirmativa lógica do slide anterior!
- Isto é, ao observar uma forte correlação, gostaríamos de concluir que uma variável causa este efeito na outra
- Infelizmente isto não é possível!
- Lembre-se: a significância do teste indica a probabilidade de se cometer um erro do tipo I (falso positivo).

Repita várias vezes mentalmente

Correlação não implica em causalidade.

Exemplo

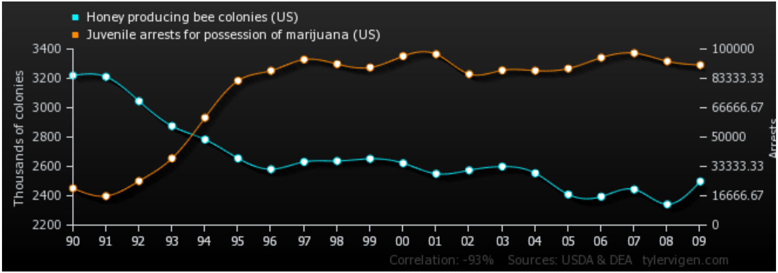
Gasto com C&T (EUA) x Suicídios por enforcamento



Correlação: 0.992082
(Fonte: Spurious correlations)

Exemplo

Produção de mel x Prisões por posse de maconha

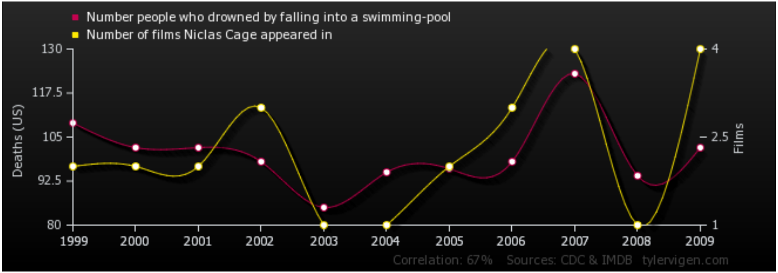


Correlação: -0.933389

(Fonte: Spurious correlations)

Exemplo

Afogamentos em piscina x Filmes com Nicholas Cage



Correlação: 0.666004
(Fonte: Spurious correlations)

Causa e efeito

Ao encontrar uma forte correlação, deve-se sempre se perguntar:

1. Há uma relação direta de causa e efeito entre as variáveis? (X causa Y?)
2. Há uma relação inversa de causa e efeito entre as variáveis? (Y causa X?)
3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável (ou mais) que não foi analisada?
4. É possível que a relação entre duas variáveis seja uma coincidência?

Causa e efeito

Ao encontrar uma forte correlação, deve-se sempre se perguntar:

1. Há uma relação direta de causa e efeito entre as variáveis? (X causa Y?)
2. Há uma relação inversa de causa e efeito entre as variáveis? (Y causa X?)
3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável (ou mais) que não foi analisada?
4. É possível que a relação entre duas variáveis seja uma coincidência?

Causa e efeito

Ao encontrar uma forte correlação, deve-se sempre se perguntar:

1. Há uma relação direta de causa e efeito entre as variáveis? (X causa Y?)
2. Há uma relação inversa de causa e efeito entre as variáveis? (Y causa X?)
3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável (ou mais) que não foi analisada?
4. É possível que a relação entre duas variáveis seja uma coincidência?

Causa e efeito

Ao encontrar uma forte correlação, deve-se sempre se perguntar:

1. Há uma relação direta de causa e efeito entre as variáveis? (X causa Y?)
2. Há uma relação inversa de causa e efeito entre as variáveis? (Y causa X?)
3. É possível que a relação entre as variáveis possa ser causada por uma terceira variável (ou mais) que não foi analisada?
4. É possível que a relação entre duas variáveis seja uma coincidência?

Resumo

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
- Qual proporção (porcentagem) da variabilidade pode ser explicada pelas variáveis analisadas?
- Quão bem a reta regressora se ajusta aos dados?

Resumo

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
- Qual proporção (porcentagem) da variabilidade pode ser explicada pelas variáveis analisadas?
- Quão bem a reta regressora se ajusta aos dados?

Resumo

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
- Qual proporção (porcentagem) da variabilidade pode ser explicada pelas variáveis analisadas?
- Quão bem a reta regressora se ajusta aos dados?

Resumo

- É necessário investigar a relação entre as variáveis!
- O que pode explicar a relação observada?
- Qual proporção (porcentagem) da variabilidade pode ser explicada pelas variáveis analisadas?
- Quão bem a reta regressora se ajusta aos dados?