Probabilidade e estatística - Aula 23

O Coeficiente de Determinação e Correlação

Dr. Giannini Italino Alves Vieira

Universidade Federal do Ceará - Campus de Crateús

2024

Coeficiente de determinação

- Recorde que ajustar um modelo de regressão linear simples requer várias suposições.
- Por exemplo:
  - A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância σ^2 constante.
  - Teste de hipóteses e intervalos de confiança requerem que os erros sejam normalmente distribuídos.
- Além disso, consideramos ainda que a ordem do modelo esteja correta, ou seja, o fenômeno se comporta de maneira linear.
- Então, essas suposições sempre devem ser verificadas e a realização de análises para examinar a adequação do modelo que se está testando deve ser feita.
- Veremos, a seguir, uma medida útil para julgar a adequação do modelo de regressão.

Coeficiente de determinação

- Recorde que ajustar um modelo de regressão linear simples requer várias suposições.
- Por exemplo:
  - A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância σ^2 constante.
  - Teste de hipóteses e intervalos de confiança requerem que os erros sejam normalmente distribuídos.
- Além disso, consideramos ainda que a ordem do modelo esteja correta, ou seja, o fenômeno se comporta de maneira linear.
- Então, essas suposições sempre devem ser verificadas e a realização de análises para examinar a adequação do modelo que se está testando deve ser feita.
- Veremos, a seguir, uma medida útil para julgar a adequação do modelo de regressão.

Coeficiente de determinação

- Recorde que ajustar um modelo de regressão linear simples requer várias suposições.
- Por exemplo:
  - A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância σ^2 constante.
  - Teste de hipóteses e intervalos de confiança requerem que os erros sejam normalmente distribuídos.
- Além disso, consideramos ainda que a ordem do modelo esteja correta, ou seja, o fenômeno se comporta de maneira linear.
- Então, essas suposições sempre devem ser verificadas e a realização de análises para examinar a adequação do modelo que se está testando deve ser feita.
- Veremos, a seguir, uma medida útil para julgar a adequação do modelo de regressão.

Coeficiente de determinação

- Recorde que ajustar um modelo de regressão linear simples requer várias suposições.
- Por exemplo:
  - A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância σ^2 constante.
  - Teste de hipóteses e intervalos de confiança requerem que os erros sejam normalmente distribuídos.
- Além disso, consideramos ainda que a ordem do modelo esteja correta, ou seja, o fenômeno se comporta de maneira linear.
- Então, essas suposições sempre devem ser verificadas e a realização de análises para examinar a adequação do modelo que se está testando deve ser feita.
- Veremos, a seguir, uma medida útil para julgar a adequação do modelo de regressão.

Coeficiente de determinação

- Recorde que ajustar um modelo de regressão linear simples requer várias suposições.
- Por exemplo:
  - A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância σ^2 constante.
  - Teste de hipóteses e intervalos de confiança requerem que os erros sejam normalmente distribuídos.
- Além disso, consideramos ainda que a ordem do modelo esteja correta, ou seja, o fenômeno se comporta de maneira linear.
- Então, essas suposições sempre devem ser verificadas e a realização de análises para examinar a adequação do modelo que se está testando deve ser feita.
- Veremos, a seguir, uma medida útil para julgar a adequação do modelo de regressão.

Coeficiente de determinação

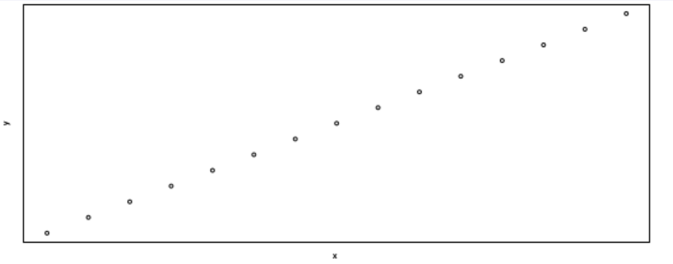
- Recorde que ajustar um modelo de regressão linear simples requer várias suposições.
- Por exemplo:
  - A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância σ^2 constante.
  - Teste de hipóteses e intervalos de confiança requerem que os erros sejam normalmente distribuídos.
- Além disso, consideramos ainda que a ordem do modelo esteja correta, ou seja, o fenômeno se comporta de maneira linear.
- Então, essas suposições sempre devem ser verificadas e a realização de análises para examinar a adequação do modelo que se está testando deve ser feita.
- Veremos, a seguir, uma medida útil para julgar a adequação do modelo de regressão.

Coeficiente de determinação

- Recorde que ajustar um modelo de regressão linear simples requer várias suposições.
- Por exemplo:
  - A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância σ^2 constante.
  - Teste de hipóteses e intervalos de confiança requerem que os erros sejam normalmente distribuídos.
- Além disso, consideramos ainda que a ordem do modelo esteja correta, ou seja, o fenômeno se comporta de maneira linear.
- Então, essas suposições sempre devem ser verificadas e a realização de análises para examinar a adequação do modelo que se está testando deve ser feita.
- Veremos, a seguir, uma medida útil para julgar a adequação do modelo de regressão.

Coeficiente de determinação- Motivação

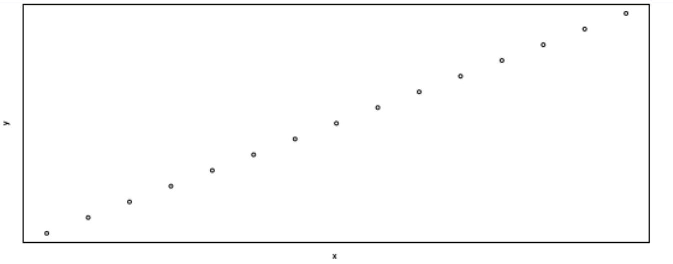
- A fim de motivar a ideia do coeficiente de determinação que veremos a seguir, considere as seguintes situações:



- Note que na figura acima todos os pontos dispõe-se exatamente em uma reta.
- Nesse caso, note que toda a variação amostral em y pode ser atribuída ao fato de x e y estarem relacionados linearmente.
- Ou seja, toda a variabilidade nos dados esta sendo explicada pelo modelo de regressão ajustado.

Coeficiente de determinação- Motivação

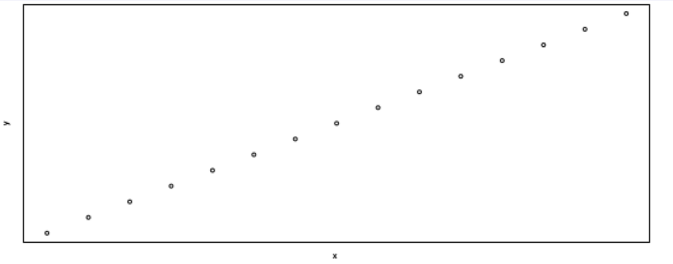
- A fim de motivar a ideia do coeficiente de determinação que veremos a seguir, considere as seguintes situações:



- Note que na figura acima todos os pontos dispõe-se exatamente em uma reta.
- Nesse caso, note que toda a variação amostral em y pode ser atribuída ao fato de x e y estarem relacionados linearmente.
- Ou seja, toda a variabilidade nos dados esta sendo explicada pelo modelo de regressão ajustado.

Coeficiente de determinação- Motivação

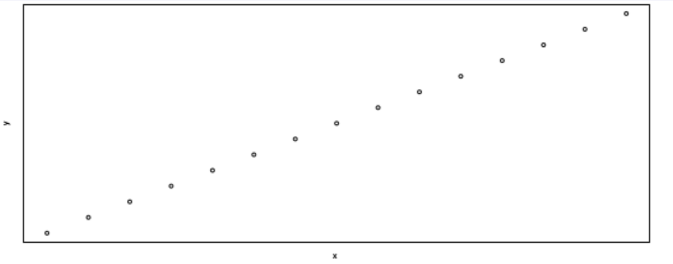
- A fim de motivar a ideia do coeficiente de determinação que veremos a seguir, considere as seguintes situações:



- Note que na figura acima todos os pontos dispõe-se exatamente em uma reta.
- Nesse caso, note que toda a variação amostral em y pode ser atribuída ao fato de x e y estarem relacionados linearmente.
- Ou seja, toda a variabilidade nos dados esta sendo explicada pelo modelo de regressão ajustado.

Coeficiente de determinação- Motivação

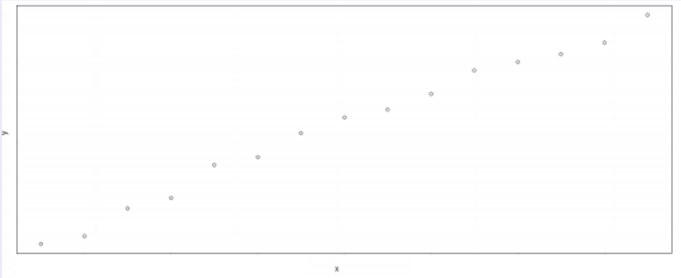
- A fim de motivar a ideia do coeficiente de determinação que veremos a seguir, considere as seguintes situações:



- Note que na figura acima todos os pontos dispõe-se exatamente em uma reta.
- Nesse caso, note que toda a variação amostral em y pode ser atribuída ao fato de x e y estarem relacionados linearmente.
- Ou seja, toda a variabilidade nos dados esta sendo explicada pelo modelo de regressão ajustado.

Coeficiente de determinação- Motivação

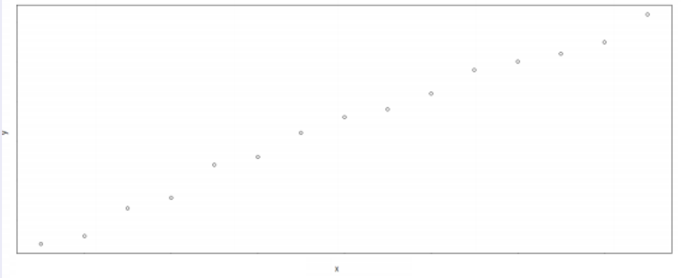
- Na situação da figura abaixo, note que os pontos não se dispõem exatamente em uma reta.



- Mas, em comparação com a variabilidade de y, os desvios da reta de mínimo quadrados são pequenos.
- Ou seja, nesse caso, temos que grande parte da variabilidade de y observada está sendo explicada pelo modelo de regressão linear simples.

Coeficiente de determinação- Motivação

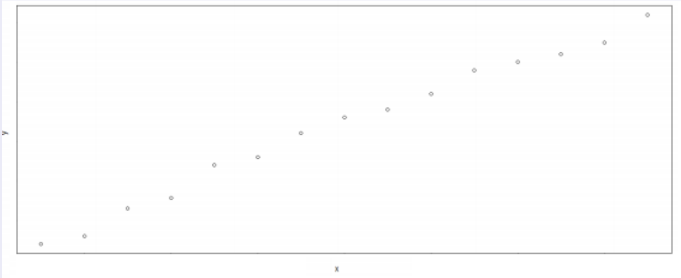
- Na situação da figura abaixo, note que os pontos não se dispõem exatamente em uma reta.



- Mas, em comparação com a variabilidade de y, os desvios da reta de mínimo quadrados são pequenos.
- Ou seja, nesse caso, temos que grande parte da variabilidade de y observada está sendo explicada pelo modelo de regressão linear simples.

Coeficiente de determinação- Motivação

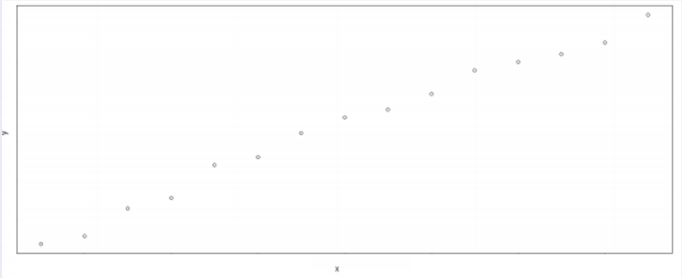
- Na situação da figura abaixo, note que os pontos não se dispõem exatamente em uma reta.



- Mas, em comparação com a variabilidade de y, os desvios da reta de mínimo quadrados são pequenos.
- Ou seja, nesse caso, temos que grande parte da variabilidade de y observada está sendo explicada pelo modelo de regressão linear simples.

Coeficiente de determinação- Motivação

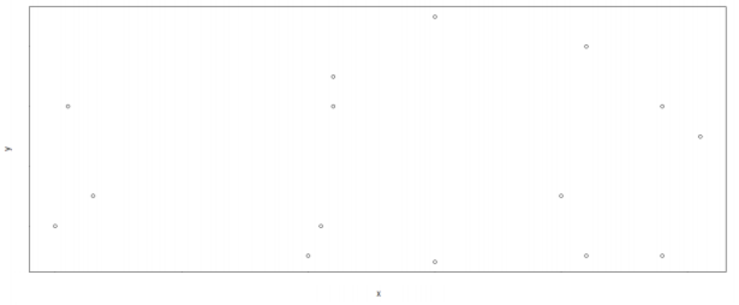
- Na situação da figura abaixo, note que os pontos não se dispõem exatamente em uma reta.



- Mas, em comparação com a variabilidade de y, os desvios da reta de mínimo quadrados são pequenos.
- Ou seja, nesse caso, temos que grande parte da variabilidade de y observada está sendo explicada pelo modelo de regressão linear simples.

Coeficiente de determinação

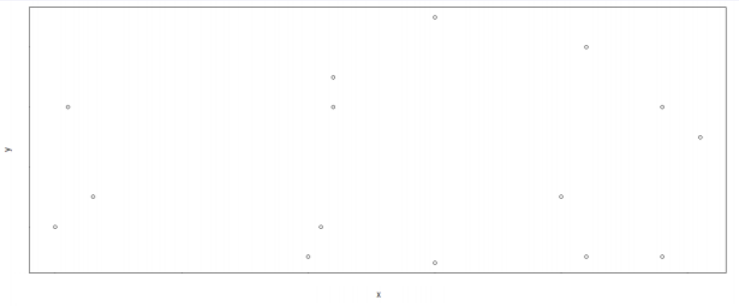
- Já na situação ilustrada na figura abaixo, note que há uma variação significativa ao redor da reta de mínimos quadrados, em relação a variação de y.



- Ou seja, nesse caso o modelo de regressão linear simples não consegue explicar a variação em y relacionando-o a x.

Coeficiente de determinação

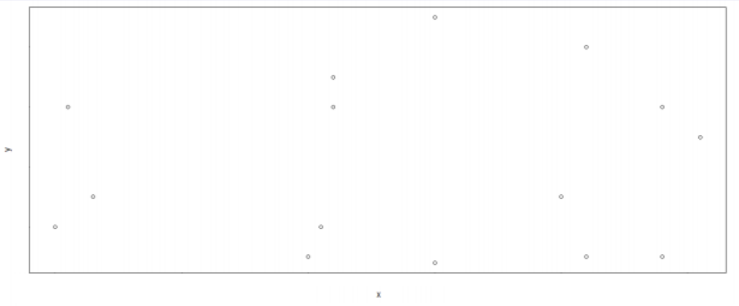
- Já na situação ilustrada na figura abaixo, note que há uma variação significativa ao redor da reta de mínimos quadrados, em relação a variação de y.



- Ou seja, nesse caso o modelo de regressão linear simples não consegue explicar a variação em y relacionando-o a x.

Coeficiente de determinação

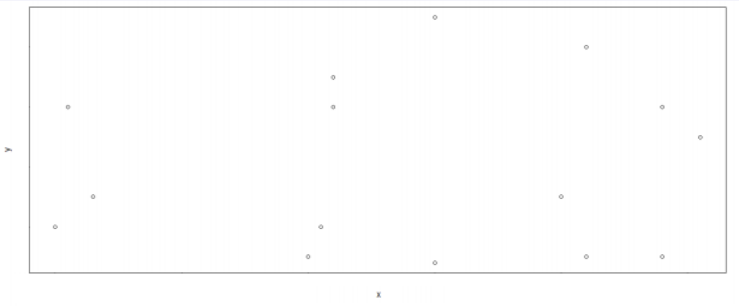
- Já na situação ilustrada na figura abaixo, note que há uma variação significativa ao redor da reta de mínimos quadrados, em relação a variação de y.



- Ou seja, nesse caso o modelo de regressão linear simples não consegue explicar a variação em y relacionando-o a x.

Coeficiente de determinação

- Já na situação ilustrada na figura abaixo, note que há uma variação significativa ao redor da reta de mínimos quadrados, em relação a variação de y.



- Ou seja, nesse caso o modelo de regressão linear simples não consegue explicar a variação em y relacionando-o a x.

Coeficiente de determinação

- Veremos agora uma medida largamente usada para julgar a adequação de um modelo de regressão linear simples.

Coeficiente de determinação

Definição: O coeficiente de determinação, denotado por R^2, é definido como a quantidade

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}$$

em que

- $SQ_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (soma dos quadrados da regressão);
- $SQ_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ (soma dos quadrados totais);
- $SQ_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (soma dos quadrados dos resíduos).

Coeficiente de determinação

- Veremos agora uma medida largamente usada para julgar a adequação de um modelo de regressão linear simples.

Coeficiente de determinação

Definição: O coeficiente de determinação, denotado por R^2, é definido como a quantidade

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}$$

em que

- $SQ_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (soma dos quadrados da regressão);
- $SQ_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ (soma dos quadrados totais);
- $SQ_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (soma dos quadrados dos resíduos).

Coeficiente de determinação

- Veremos agora uma medida largamente usada para julgar a adequação de um modelo de regressão linear simples.

Coeficiente de determinação

Definição: O coeficiente de determinação, denotado por R^2, é definido como a quantidade

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}$$

em que

- $SQ_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ (soma dos quadrados da regressão);
- $SQ_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ (soma dos quadrados totais);
- $SQ_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ (soma dos quadrados dos resíduos).

Coeficiente de determinação

- Veremos agora uma medida largamente usada para julgar a adequação de um modelo de regressão linear simples.

Coeficiente de determinação

Definição: O coeficiente de determinação, denotado por R^2, é definido como a quantidade

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}$$

em que

- $SQ_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ (soma dos quadrados da regressão);
- $SQ_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ (soma dos quadrados totais);
- $SQ_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ (soma dos quadrados dos resíduos).

Coeficiente de determinação

- Veremos agora uma medida largamente usada para julgar a adequação de um modelo de regressão linear simples.

Coeficiente de determinação

Definição: O coeficiente de determinação, denotado por R^2, é definido como a quantidade

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}$$

em que

- $SQ_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ (soma dos quadrados da regressão);
- $SQ_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ (soma dos quadrados totais);
- $SQ_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ (soma dos quadrados dos resíduos).

Coeficiente de determinação

- Veremos agora uma medida largamente usada para julgar a adequação de um modelo de regressão linear simples.

Coeficiente de determinação

Definição: O coeficiente de determinação, denotado por R^2, é definido como a quantidade

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}$$

em que

- $SQ_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (soma dos quadrados da regressão);
- $SQ_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ (soma dos quadrados totais);
- $SQ_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (soma dos quadrados dos resíduos).

Coeficiente de determinação

- R^2 é frequentemente usado para julgar a adequação de um modelo de regressão.
- Esse coeficiente é interpretado como a proporção de variabilidade de y observada que pode ser explicada pelo modelo de regressão linear simples.
- É possível mostrar que $SQ_T = SQ_R + SQ_E$ (identidade de análise de variância).
- Logo, usando esse resultado temos que $0 \leq R^2 \leq 1$.
- Quanto maior for o valor de R^2 (mais próximo de 1), mais o modelo de regressão linear consegue explicar a variabilidade de y.
- Quando R^2 for pequeno, então em geral o analista vai ter que procurar um modelo alternativo, como por exemplo, um modelo não linear ou um modelo de regressão linear múltipla, com mais de uma variável independente, de maneira a explicar mais eficientemente a variação de y.

Coeficiente de determinação

- R^2 é frequentemente usado para julgar a adequação de um modelo de regressão.
- Esse coeficiente é interpretado como a proporção de variabilidade de y observada que pode ser explicada pelo modelo de regressão linear simples.
- É possível mostrar que $SQ_T = SQ_R + SQ_E$ (identidade de análise de variância).
- Logo, usando esse resultado temos que $0 \leq R^2 \leq 1$.
- Quanto maior for o valor de R^2 (mais próximo de 1), mais o modelo de regressão linear consegue explicar a variabilidade de y.
- Quando R^2 for pequeno, então em geral o analista vai ter que procurar um modelo alternativo, como por exemplo, um modelo não linear ou um modelo de regressão linear múltipla, com mais de uma variável independente, de maneira a explicar mais eficientemente a variação de y.

Coeficiente de determinação

- R^2 é frequentemente usado para julgar a adequação de um modelo de regressão.
- Esse coeficiente é interpretado como a proporção de variabilidade de y observada que pode ser explicada pelo modelo de regressão linear simples.
- É possível mostrar que $SQ_T = SQ_R + SQ_E$ (identidade de análise de variância).
- Logo, usando esse resultado temos que $0 \leq R^2 \leq 1$.
- Quanto maior for o valor de R^2 (mais próximo de 1), mais o modelo de regressão linear consegue explicar a variabilidade de y.
- Quando R^2 for pequeno, então em geral o analista vai ter que procurar um modelo alternativo, como por exemplo, um modelo não linear ou um modelo de regressão linear múltipla, com mais de uma variável independente, de maneira a explicar mais eficientemente a variação de y.

Coeficiente de determinação

- R^2 é frequentemente usado para julgar a adequação de um modelo de regressão.
- Esse coeficiente é interpretado como a proporção de variabilidade de y observada que pode ser explicada pelo modelo de regressão linear simples.
- É possível mostrar que $SQ_T = SQ_R + SQ_E$ (identidade de análise de variância).
- Logo, usando esse resultado temos que $0 \leq R^2 \leq 1$.
- Quanto maior for o valor de R^2 (mais próximo de 1), mais o modelo de regressão linear consegue explicar a variabilidade de y.
- Quando R^2 for pequeno, então em geral o analista vai ter que procurar um modelo alternativo, como por exemplo, um modelo não linear ou um modelo de regressão linear múltipla, com mais de uma variável independente, de maneira a explicar mais eficientemente a variação de y.

Coeficiente de determinação

- R^2 é frequentemente usado para julgar a adequação de um modelo de regressão.
- Esse coeficiente é interpretado como a proporção de variabilidade de y observada que pode ser explicada pelo modelo de regressão linear simples.
- É possível mostrar que $SQ_T = SQ_R + SQ_E$ (identidade de análise de variância).
- Logo, usando esse resultado temos que $0 \leq R^2 \leq 1$.
- Quanto maior for o valor de R^2 (mais próximo de 1), mais o modelo de regressão linear consegue explicar a variabilidade de y.
- Quando R^2 for pequeno, então em geral o analista vai ter que procurar um modelo alternativo, como por exemplo, um modelo não linear ou um modelo de regressão linear múltipla, com mais de uma variável independente, de maneira a explicar mais eficientemente a variação de y.

Coeficiente de determinação

- R^2 é frequentemente usado para julgar a adequação de um modelo de regressão.
- Esse coeficiente é interpretado como a proporção de variabilidade de y observada que pode ser explicada pelo modelo de regressão linear simples.
- É possível mostrar que $SQ_T = SQ_R + SQ_E$ (identidade de análise de variância).
- Logo, usando esse resultado temos que $0 \leq R^2 \leq 1$.
- Quanto maior for o valor de R^2 (mais próximo de 1), mais o modelo de regressão linear consegue explicar a variabilidade de y.
- Quando R^2 for pequeno, então em geral o analista vai ter que procurar um modelo alternativo, como por exemplo, um modelo não linear ou um modelo de regressão linear múltipla, com mais de uma variável independente, de maneira a explicar mais eficientemente a variação de y.

Exemplo

- Por exemplo, vamos considerar novamente o exemplo visto na aula anterior, ou seja:

Exemplo: Um artigo na revista *IEEE transactions on instrumentation and measurement ["Direct, fast, and accurate measurement of V_T and K of an MOS transistor using a V_T-sift circuit" (Vol. 40, 1991, pp. 951-955)]* descreveu o uso de um modelo de regressão linear simples para expressar a corrente y (em miliamperes), como função da diferença de voltagem x (em volts). Os dados são fornecidos a seguir:

x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
y	0.734	0.886	1.04	1.19	1.35	1.5	1.66	1.81	1.97	2.12

- Vamos agora obter o coeficiente de determinação do modelo de regressão que obtivemos.
- Recorde que o modelo ajustado foi $\hat{y}_i = -0.966824 + 1.54376x_i$.

Exemplo

- Por exemplo, vamos considerar novamente o exemplo visto na aula anterior, ou seja:

Exemplo: Um artigo na revista *IEEE transactions on instrumentation and measurement ["Direct, fast, and accurate measurement of* V_T *and K of an MOS transistor using a* V_T*-sift circuit" (Vol. 40, 1991, pp. 951-955)]* descreveu o uso de um modelo de regressão linear simples para expressar a corrente y (em miliamperes), como função da diferença de voltagem x (em volts). Os dados são fornecidos a seguir:

x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
y	0.734	0.886	1.04	1.19	1.35	1.5	1.66	1.81	1.97	2.12

- Vamos agora obter o coeficiente de determinação do modelo de regressão que obtivemos.
- Recorde que o modelo ajustado foi $\hat{y}_i = -0.966824 + 1.54376x_i$.

Exemplo

- Por exemplo, vamos considerar novamente o exemplo visto na aula anterior, ou seja:

Exemplo: Um artigo na revista *IEEE transactions on instrumentation and measurement ["Direct, fast, and accurate measurement of* V_T *and K of an MOS transistor using a* V_T*-sift circuit" (Vol. 40, 1991, pp. 951-955)]* descreveu o uso de um modelo de regressão linear simples para expressar a corrente y (em miliamperes), como função da diferença de voltagem x (em volts). Os dados são fornecidos a seguir:

x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
y	0.734	0.886	1.04	1.19	1.35	1.5	1.66	1.81	1.97	2.12

- Vamos agora obter o coeficiente de determinação do modelo de regressão que obtivemos.
- Recorde que o modelo ajustado foi $\hat{y}_i = -0.966824 + 1.54376x_i$.

Exemplo

- Por exemplo, vamos considerar novamente o exemplo visto na aula anterior, ou seja:

Exemplo: Um artigo na revista *IEEE transactions on instrumentation and measurement ["Direct, fast, and accurate measurement of V_T and K of an MOS transistor using a V_T-sift circuit" (Vol. 40, 1991, pp. 951-955)]* descreveu o uso de um modelo de regressão linear simples para expressar a corrente y (em miliamperes), como função da diferença de voltagem x (em volts). Os dados são fornecidos a seguir:

x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
y	0.734	0.886	1.04	1.19	1.35	1.5	1.66	1.81	1.97	2.12

- Vamos agora obter o coeficiente de determinação do modelo de regressão que obtivemos.
- Recorde que o modelo ajustado foi $\hat{y}_i = -0.966824 + 1.54376x_i$.

Exemplo

- Por exemplo, vamos considerar novamente o exemplo visto na aula anterior, ou seja:

Exemplo: Um artigo na revista *IEEE transactions on instrumentation and measurement ["Direct, fast, and accurate measurement of V_T and K of an MOS transistor using a V_T-sift circuit" (Vol. 40, 1991, pp. 951-955)]* descreveu o uso de um modelo de regressão linear simples para expressar a corrente y (em miliamperes), como função da diferença de voltagem x (em volts). Os dados são fornecidos a seguir:

x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
y	0.734	0.886	1.04	1.19	1.35	1.5	1.66	1.81	1.97	2.12

- Vamos agora obter o coeficiente de determinação do modelo de regressão que obtivemos.
- Recorde que o modelo ajustado foi $\hat{y}_i = -0.966824 + 1.54376x_i$.

- Recorde que da aula passada obtivemos que:

$$SQ_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 22.301 - 10(1.426)^2 = 1.96624$$

- E sabemos que

$$SQ_E = SQ_T - \hat{\beta}_1 S_{xy} = 1.96624 - 1.54376(1.2736) \approx 0.000107264.$$

- Logo, temos que

$$R^2 = 1 - \frac{SQ_E}{SQ_T} = 1 - \frac{0.000107264}{1.96624} \approx 0.999945.$$

- Ou seja, aproximadamente 99.99% da variação observada da corrente (y) é atribuída (pode ser explicada) por a relação linear aproximada entre corrente (y) e diferença de voltagem (x).

- Recorde que da aula passada obtivemos que:

$$SQ_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 22.301 - 10(1.426)^2 = 1.96624$$

- E sabemos que

$$SQ_E = SQ_T - \hat{\beta}_1 S_{xy} = 1.96624 - 1.54376(1.2736) \approx 0.000107264.$$

- Logo, temos que

$$R^2 = 1 - \frac{SQ_E}{SQ_T} = 1 - \frac{0.000107264}{1.96624} \approx 0.999945.$$

- Ou seja, aproximadamente 99.99% da variação observada da corrente (y) é atribuída (pode ser explicada) por a relação linear aproximada entre corrente (y) e diferença de voltagem (x).

- Recorde que da aula passada obtivemos que:

$$SQ_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 22.301 - 10(1.426)^2 = 1.96624$$

- E sabemos que

$$SQ_E = SQ_T - \hat{\beta}_1 S_{xy} = 1.96624 - 1.54376(1.2736) \approx 0.000107264.$$

- Logo, temos que

$$R^2 = 1 - \frac{SQ_E}{SQ_T} = 1 - \frac{0.000107264}{1.96624} \approx 0.999945.$$

- Ou seja, aproximadamente 99.99% da variação observada da corrente (y) é atribuída (pode ser explicada) por a relação linear aproximada entre corrente (y) e diferença de voltagem (x).

- Recorde que da aula passada obtivemos que:

$$SQ_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 22.301 - 10(1.426)^2 = 1.96624$$

- E sabemos que

$$SQ_E = SQ_T - \hat{\beta}_1 S_{xy} = 1.96624 - 1.54376(1.2736) \approx 0.000107264.$$

- Logo, temos que

$$R^2 = 1 - \frac{SQ_E}{SQ_T} = 1 - \frac{0.000107264}{1.96624} \approx 0.999945.$$

- Ou seja, aproximadamente 99.99% da variação observada da corrente (y) é atribuída (pode ser explicada) por a relação linear aproximada entre corrente (y) e diferença de voltagem (x).

- Recorde que da aula passada obtivemos que:

$$SQ_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 22.301 - 10(1.426)^2 = 1.96624$$

- E sabemos que

$$SQ_E = SQ_T - \hat{\beta}_1 S_{xy} = 1.96624 - 1.54376(1.2736) \approx 0.000107264.$$

- Logo, temos que

$$R^2 = 1 - \frac{SQ_E}{SQ_T} = 1 - \frac{0.000107264}{1.96624} \approx 0.999945.$$

- Ou seja, aproximadamente 99.99% da variação observada da corrente (y) é atribuída (pode ser explicada) por a relação linear aproximada entre corrente (y) e diferença de voltagem (x).

Coeficiente de correlação amostral

- Existem muitas situações em que o objetivo de estudar o comportamento conjunto de duas variáveis é verificar se elas estão relacionada e não, necessariamente, usar uma para predizer o valor da outra.
- Veremos agora uma medida que especifica o quão fortemente duas variáveis x e y estão relacionadas, em uma amostra.
- Essa medida é chamada de coeficiente de correlação amostral e mede o grau de relação linear entre as variáveis.

Coeficiente de correlação amostral

- Existem muitas situações em que o objetivo de estudar o comportamento conjunto de duas variáveis é verificar se elas estão relacionada e não, necessariamente, usar uma para predizer o valor da outra.
- Veremos agora uma medida que especifica o quão fortemente duas variáveis x e y estão relacionadas, em uma amostra.
- Essa medida é chamada de coeficiente de correlação amostral e mede o grau de relação linear entre as variáveis.

Coeficiente de correlação amostral

- Existem muitas situações em que o objetivo de estudar o comportamento conjunto de duas variáveis é verificar se elas estão relacionada e não, necessariamente, usar uma para predizer o valor da outra.
- Veremos agora uma medida que especifica o quão fortemente duas variáveis x e y estão relacionadas, em uma amostra.
- Essa medida é chamada de coeficiente de correlação amostral e mede o grau de relação linear entre as variáveis.

Coeficiente de correlação amostral

Definição: O coeficiente de correlação amostral de n pares $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ é definido por

$$R = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Propriedades mais importantes sobre o coeficiente de correlação amostral

- O valor de R não depende de qual das duas variáveis em estudo é chamada de x ou de y.
- O valor de R independe das unidades com os quais x e y são medidas.
- $-1 \leq R \leq 1$;
- $R = 1$ se, e somente se, todos os pares (x_i, y_i) estiverem alinhados em linha reta com coeficiente angular positivo. De maneira análoga, $R = -1$ se, e somente se, todos os pares (x_i, y_i) estiverem alinhados em linha reta com coeficiente angular negativo.
- O quadrado do coeficiente de correlação amostral fornece o valor do coeficiente de determinação que resultaria de um ajuste do modelo de regressão linear simples.
- Um valor de R próximo de zero indica uma falta de relação linear.

Propriedades mais importantes sobre o coeficiente de correlação amostral

- O valor de R não depende de qual das duas variáveis em estudo é chamada de x ou de y.
- O valor de R independe das unidades com os quais x e y são medidas.
- $-1 \leq R \leq 1$;
- $R = 1$ se, e somente se, todos os pares (x_i, y_i) estiverem alinhados em linha reta com coeficiente angular positivo. De maneira análoga, $R = -1$ se, e somente se, todos os pares (x_i, y_i) estiverem alinhados em linha reta com coeficiente angular negativo.
- O quadrado do coeficiente de correlação amostral fornece o valor do coeficiente de determinação que resultaria de um ajuste do modelo de regressão linear simples.
- Um valor de R próximo de zero indica uma falta de relação linear.

Propriedades mais importantes sobre o coeficiente de correlação amostral

- O valor de R não depende de qual das duas variáveis em estudo é chamada de x ou de y.
- O valor de R independe das unidades com os quais x e y são medidas.
- $-1 \leq R \leq 1$;
- $R = 1$ se, e somente se, todos os pares (x_i, y_i) estiverem alinhados em linha reta com coeficiente angular positivo. De maneira análoga, $R = -1$ se, e somente se, todos os pares (x_i, y_i) estiverem alinhados em linha reta com coeficiente angular negativo.
- O quadrado do coeficiente de correlação amostral fornece o valor do coeficiente de determinação que resultaria de um ajuste do modelo de regressão linear simples.
- Um valor de R próximo de zero indica uma falta de relação linear.

Propriedades mais importantes sobre o coeficiente de correlação amostral

- O valor de R não depende de qual das duas variáveis em estudo é chamada de x ou de y.
- O valor de R independe das unidades com os quais x e y são medidas.
- $-1 \leq R \leq 1$;
- $R = 1$ se, e somente se, todos os pares (x_i, y_i) estiverem alinhados em linha reta com coeficiente angular positivo. De maneira análoga, $R = -1$ se, e somente se, todos os pares (x_i, y_i) estiverem alinhados em linha reta com coeficiente angular negativo.
- O quadrado do coeficiente de correlação amostral fornece o valor do coeficiente de determinação que resultaria de um ajuste do modelo de regressão linear simples.
- Um valor de R próximo de zero indica uma falta de relação linear.

Algumas configurações de pontos associados a diferentes valores de R

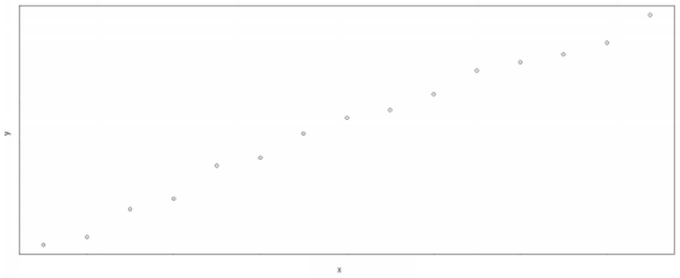


Figure 1: Correlação próximo de 1

Algumas configurações de pontos associados a diferentes valores de R

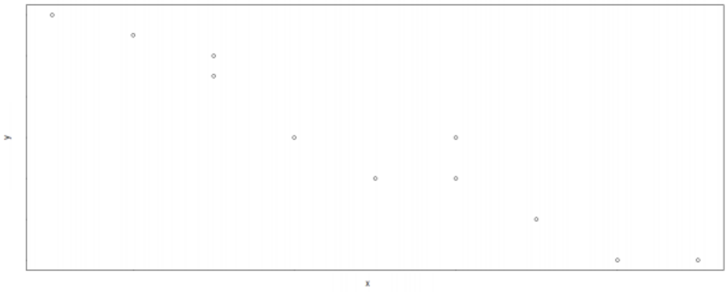


Figure 2: Correlação próximo de -1

Algumas configurações de pontos associados a diferentes valores de R

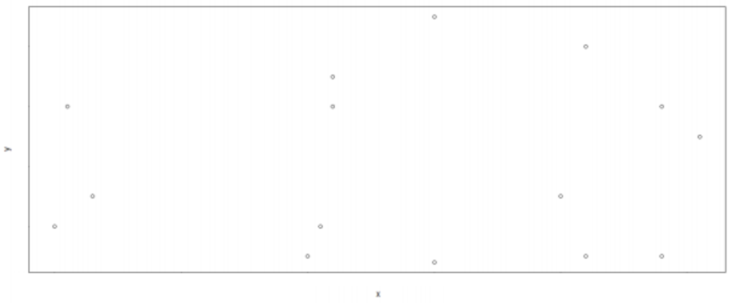


Figure 3: Correlação próximo de 0, nenhuma relação aparente.

Algumas configurações de pontos associados a diferentes valores de R

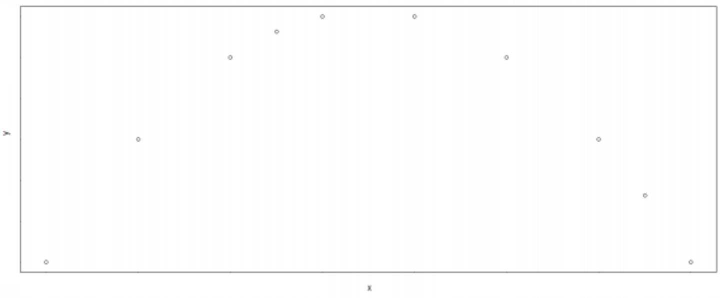


Figure 4: Correlação próximo de 0, nenhuma relação linear.

Quando é possível dizer que existe uma correlação forte entre as variáveis?

- Uma regra prática razoável é considerar que a correlação é fraca se

$$0 \leq |R| \leq 0.5$$

- A correlação e considerada moderada se

$$0.5 < |R| < 0.8$$

- E a correlação é considerada como forte se

$$0.8 \leq |R| \leq 1$$

- Note, por exemplo, que uma correlação $R = 0.5$ é considerada fraca pois implica que $R^2 = 0.25$, ou seja, significa que, em uma regressão de y em relação x, apenas 25% da variação y observada seria explicada pelo modelo.

Quando é possível dizer que existe uma correlação forte entre as variáveis?

- Uma regra prática razoável é considerar que a correlação é fraca se

$$0 \leq |R| \leq 0.5$$

- A correlação e considerada moderada se

$$0.5 < |R| < 0.8$$

- E a correlação é considerada como forte se

$$0.8 \leq |R| \leq 1$$

- Note, por exemplo, que uma correlação $R = 0.5$ é considerada fraca pois implica que $R^2 = 0.25$, ou seja, significa que, em uma regressão de y em relação x, apenas 25% da variação y observada seria explicada pelo modelo.

Quando é possível dizer que existe uma correlação forte entre as variáveis?

- Uma regra prática razoável é considerar que a correlação é fraca se

$$0 \leq |R| \leq 0.5$$

- A correlação e considerada moderada se

$$0.5 < |R| < 0.8$$

- E a correlação é considerada como forte se

$$0.8 \leq |R| \leq 1$$

- Note, por exemplo, que uma correlação $R = 0.5$ é considerada fraca pois implica que $R^2 = 0.25$, ou seja, significa que, em uma regressão de y em relação x, apenas 25% da variação y observada seria explicada pelo modelo.

Quando é possível dizer que existe uma correlação forte entre as variáveis?

- Uma regra prática razoável é considerar que a correlação é fraca se

$$0 \leq |R| \leq 0.5$$

- A correlação e considerada moderada se

$$0.5 < |R| < 0.8$$

- E a correlação é considerada como forte se

$$0.8 \leq |R| \leq 1$$

- Note, por exemplo, que uma correlação $R = 0.5$ é considerada fraca pois implica que $R^2 = 0.25$, ou seja, significa que, em uma regressão de y em relação x, apenas 25% da variação y observada seria explicada pelo modelo.

Quando é possível dizer que existe uma correlação forte entre as variáveis?

- Uma regra prática razoável é considerar que a correlação é fraca se

$$0 \leq |R| \leq 0.5$$

- A correlação e considerada moderada se

$$0.5 < |R| < 0.8$$

- E a correlação é considerada como forte se

$$0.8 \leq |R| \leq 1$$

- Note, por exemplo, que uma correlação $R = 0.5$ é considerada fraca pois implica que $R^2 = 0.25$, ou seja, significa que, em uma regressão de y em relação x, apenas 25% da variação y observada seria explicada pelo modelo.

Quando é possível dizer que existe uma correlação forte entre as variáveis?

- Uma regra prática razoável é considerar que a correlação é fraca se

$$0 \leq |R| \leq 0.5$$

- A correlação e considerada moderada se

$$0.5 < |R| < 0.8$$

- E a correlação é considerada como forte se

$$0.8 \leq |R| \leq 1$$

- Note, por exemplo, que uma correlação $R = 0.5$ é considerada fraca pois implica que $R^2 = 0.25$, ou seja, significa que, em uma regressão de y em relação x, apenas 25% da variação y observada seria explicada pelo modelo.

Exemplo

- Por exemplo, considerando o exemplo anterior, temos que a correlação é dada por

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- Mas recorde que já obtivemos as quantidades:
  - $S_{xy} = 1.2736,$
  - $S_{yy} = SQ_T = 1.96624$
  - $S_{xx} = 0.825.$
- Logo, substituindo em R, temos que o coeficiente de correlação é dado por:

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{1.2736}{\sqrt{0.825}\sqrt{1.96624}} \approx 0.999972.$$

Exemplo

- Por exemplo, considerando o exemplo anterior, temos que a correlação é dada por

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- Mas recorde que já obtivemos as quantidades:
  - $S_{xy} = 1.2736$,
  - $S_{yy} = SQ_T = 1.96624$
  - $S_{xx} = 0.825$.
- Logo, substituindo em R, temos que o coeficiente de correlação é dado por:

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{1.2736}{\sqrt{0.825}\sqrt{1.96624}} \approx 0.999972.$$

Exemplo

- Por exemplo, considerando o exemplo anterior, temos que a correlação é dada por

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- Mas recorde que já obtivemos as quantidades:
  - $S_{xy} = 1.2736$,
  - $S_{yy} = SQ_T = 1.96624$
  - $S_{xx} = 0.825$.
- Logo, substituindo em R, temos que o coeficiente de correlação é dado por:

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{1.2736}{\sqrt{0.825}\sqrt{1.96624}} \approx 0.999972.$$

Exemplo

- Por exemplo, considerando o exemplo anterior, temos que a correlação é dada por

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- Mas recorde que já obtivemos as quantidades:
  - $S_{xy} = 1.2736$,
  - $S_{yy} = SQ_T = 1.96624$
  - $S_{xx} = 0.825$.
- Logo, substituindo em R, temos que o coeficiente de correlação é dado por:

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{1.2736}{\sqrt{0.825}\sqrt{1.96624}} \approx 0.999972.$$

Exemplo

- Por exemplo, considerando o exemplo anterior, temos que a correlação é dada por

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- Mas recorde que já obtivemos as quantidades:
    - $S_{xy} = 1.2736$,
    - $S_{yy} = SQ_T = 1.96624$
    - $S_{xx} = 0.825$.
- Logo, substituindo em R, temos que o coeficiente de correlação é dado por:

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{1.2736}{\sqrt{0.825}\sqrt{1.96624}} \approx 0.999972.$$

Exemplo

- Por exemplo, considerando o exemplo anterior, temos que a correlação é dada por

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- Mas recorde que já obtivemos as quantidades:
  - $S_{xy} = 1.2736,$
  - $S_{yy} = SQ_T = 1.96624$
  - $S_{xx} = 0.825.$
- Logo, substituindo em R, temos que o coeficiente de correlação é dado por:

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{1.2736}{\sqrt{0.825}\sqrt{1.96624}} \approx 0.999972.$$

Exemplo

- Por exemplo, considerando o exemplo anterior, temos que a correlação é dada por

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- Mas recorde que já obtivemos as quantidades:
  - $S_{xy} = 1.2736,$
  - $S_{yy} = SQ_T = 1.96624$
  - $S_{xx} = 0.825.$
- Logo, substituindo em R, temos que o coeficiente de correlação é dado por:

$$R = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{1.2736}{\sqrt{0.825}\sqrt{1.96624}} \approx 0.999972.$$