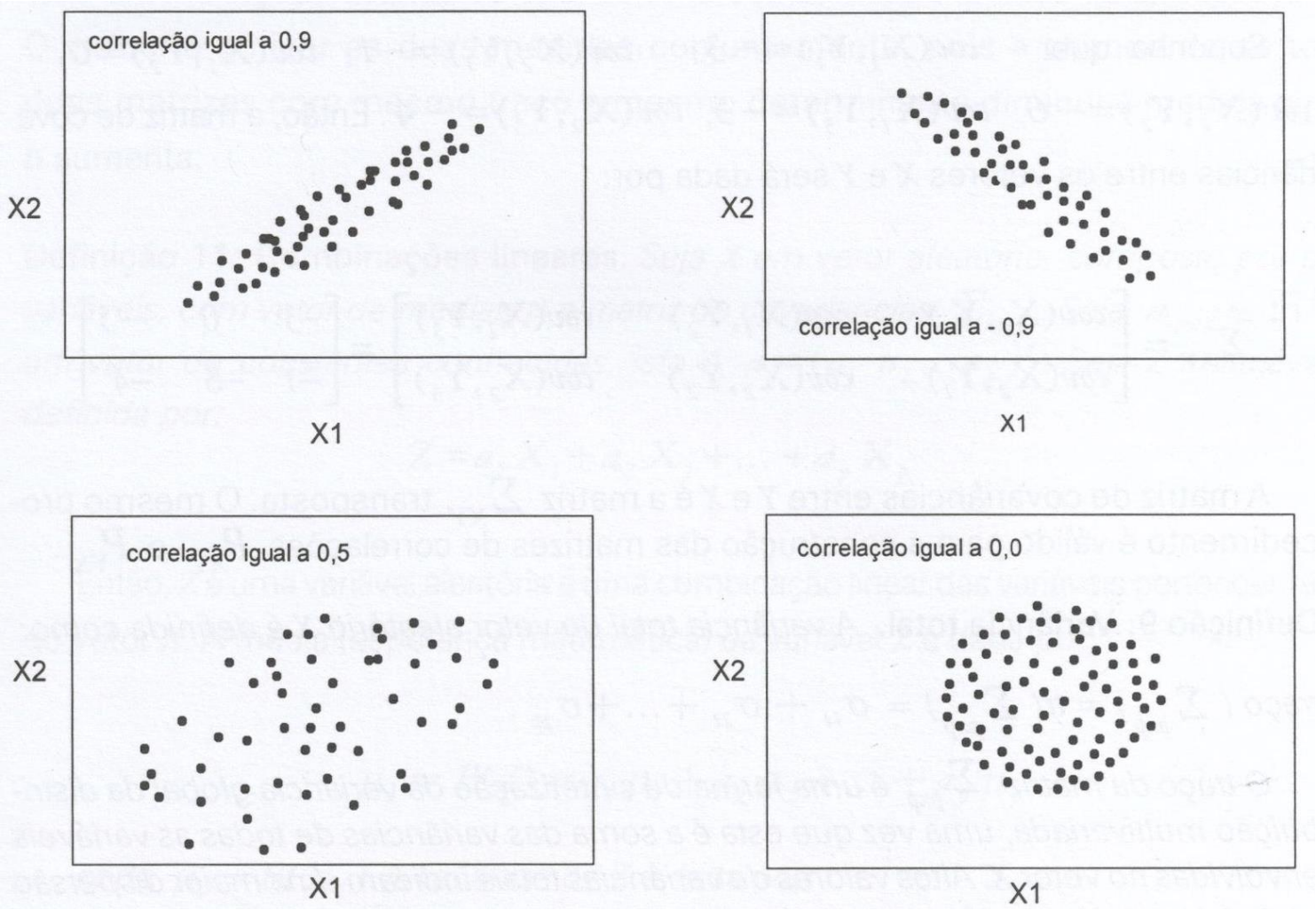
# REGRESSÃO

Prof. André Backes | @progdescomplicada

# Análise de Correlação

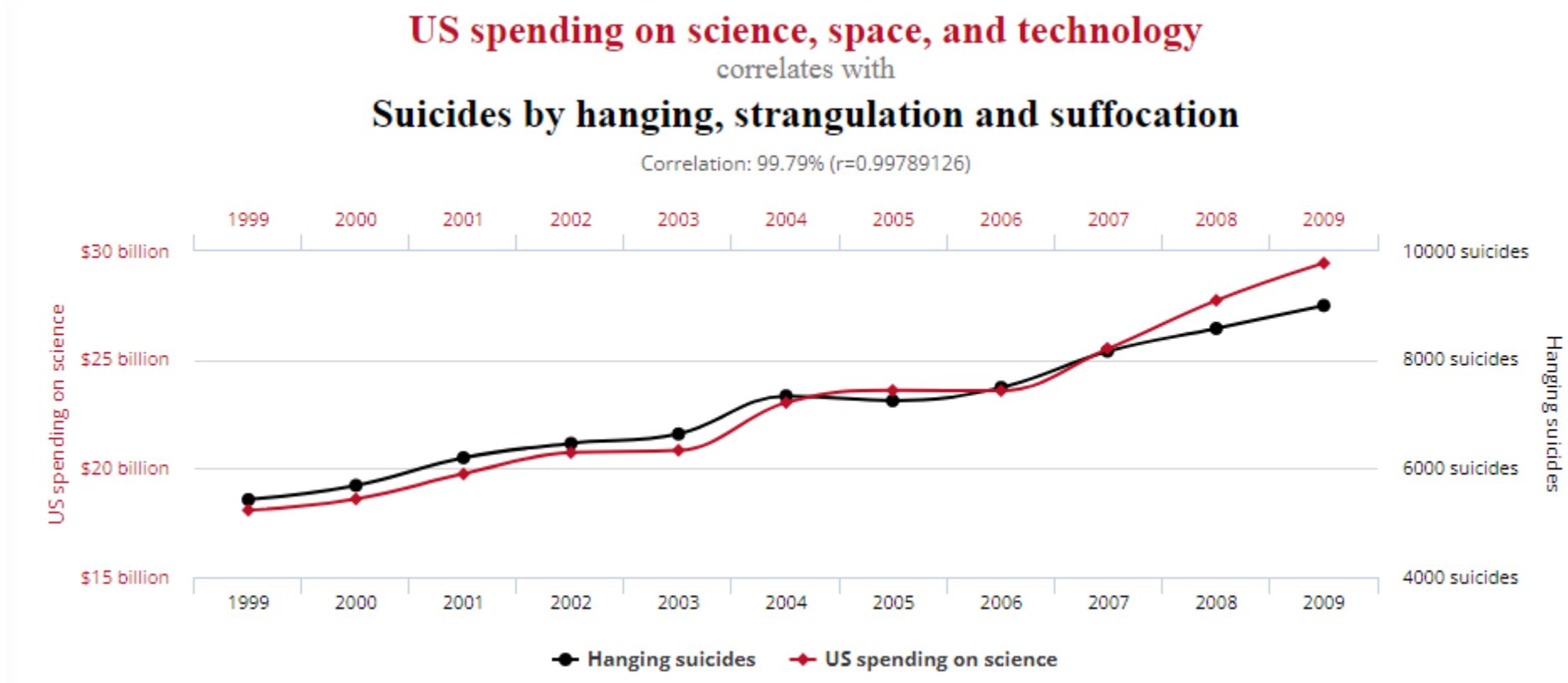
- Correlação
  - Indica a força e a direção do relacionamento linear entre dois atributos
  - Trata-se de uma medida da relação entre dois atributos, embora correlação não implique causalidade
    - Duas variáveis podem estar altamente correlacionadas e não existir relação de causa e efeito entre elas

# Análise de Correlação



# Análise de Correlação

- Correlação não implica causalidade!




**https://www.tylervigen.com/spurious-correlations**

# Análise de Correlação

- Correlação
  - Em muitas aplicações duas ou mais variáveis estão relacionadas, sendo necessário explorar a natureza desta relação
    - Correlação muito próximo de 1, ou de ( – 1), existe uma relação linear entre os dois atributos
    - Ela permite verificar se é possível ajustar um modelo que expresse a mencionada relação
    - Esse é o objetivo da **análise de regressão**

# Análise de Regressão

- O que é?
  - É uma série de técnicas voltadas para a modelagem e a investigação de relações entre dois ou mais atributos (variáveis aleatórias)
  - Exemplo
    - Na análise de correlação linear, o objetivo é determinar o grau de relacionamento entre duas variáveis.
    - Já na análise de regressão linear, o objetivo é determinar o modelo que expressa esta relação (**equação de regressão**), a qual é ajustada aos dados

# Análise de Regressão

- Para que serve?
  - Ela permite construir um modelo matemático que represente dois atributos ***x*** e ***y***
    - ***y = f(x)***, onde ***f(·)*** é a função que relaciona ***x*** e ***y***
    - ***x*** é a variável independente da equação
    - ***y = f(x)*** é a variável dependente das variações de ***x***

# Análise de Regressão

- Para que serve?
  - Podemos usar esse modelo para predizer o valor de ***y*** para um dado valor de ***x***
    - Realizar previsões sobre o comportamento futuro de algum fenômeno da realidade.
    - Neste caso extrapola-se para o futuro as relações de causa-efeito – já observadas no passado – entre as variáveis.

# Análise de Regressão

- Qual função usar?
  - Na maioria dos casos, ***f(·)*** é desconhecida
  - Cabe ao usuário escolher uma função apropriada para aproximar ***f(·)***
    - Normalmente usa-se um modelo polinomial
    - Também podemos usar o modelo para fins de otimização

# Análise de Regressão

- A análise de regressão compreende quatro tipos básicos de modelos
  - Linear simples
  - Linear multivariado
  - Não linear simples
  - Não linear multivariado

# Análise de Regressão

## Regressão simples

- Nesse tipo de regressão existe apenas uma variável de saída (***y***) e uma de entrada (***x***)
  - Exemplo: $y = f(x)$

## Regressão múltipla

- Nesse tipo de regressão existe apenas uma variável de saída (***y***) e várias de entrada ($\boldsymbol{x_i}$, ***i=1,...p***)
  - Exemplo: $y = f(x_1, x_2, \dots, x_p)$

# Análise de Regressão

## Regressão linear

- Tem esse nome porque se considera que a relação da entre as variáveis é descrita por uma função linear (equação da reta ou do plano)
  - Exemplo: $y = \alpha + \beta x$

## Regressão não linear

- Nesse caso, a relação entre as variáveis não pode ser descrita por uma função linear. Pode ser uma função exponencial ou logarítmica
  - Exemplo: $y = \alpha e^{\beta x}$

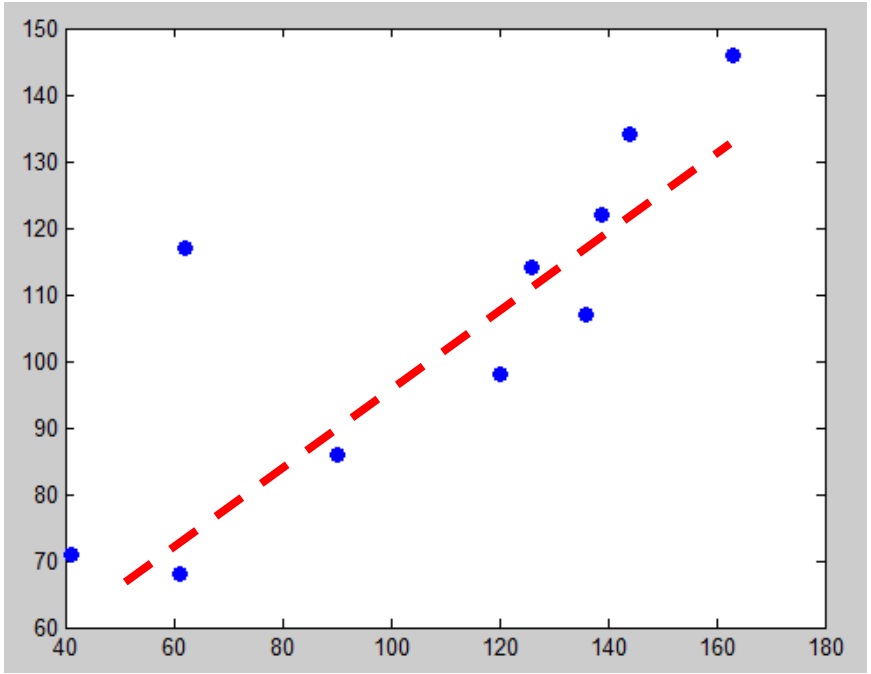
# Gráfico de dispersão (scatterplot)

- É uma representação puramente visual dos dados
  - Gráfico cartesiano dos pares de informação ***x*** e ***y*** referente a cada observação
  - Consiste de uma “nuvem” de pontos que, por sua vez, define um eixo ou direção que caracterizará o padrão de relacionamento entre as variáveis ***x*** e ***y***

# Gráfico de dispersão (scatterplot)

- A regressão será linear se observada uma tendência ou eixo linear na nuvem de pontos
  - Sempre verificar o gráfico de dispersão para saber que modelo usar

| y | x |
|---|---|
| 122 | 139 |
| 114 | 126 |
| 86 | 90 |
| 134 | 144 |
| 146 | 163 |
| 107 | 136 |
| 68 | 61 |
| 117 | 62 |
| 71 | 41 |
| 98 | 120 |

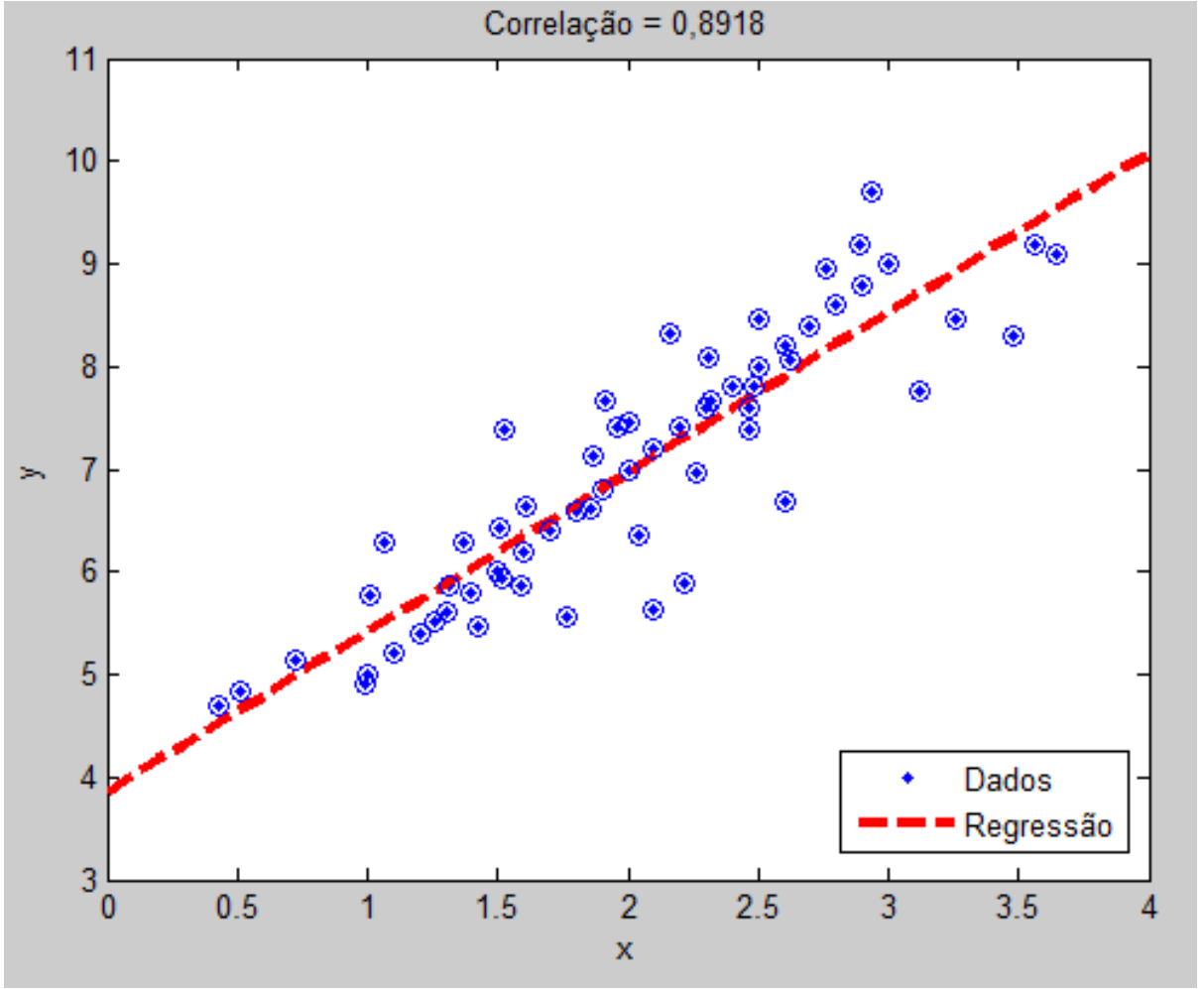


# Regressão Linear Simples

- Definições básicas
  - Existe uma única variável de saída, ***y***
    - Variável dependente
  - Existe uma (***x***) de entrada
    - variável independente ou regressora
  - Assume-se que as variáveis de entrada são medidas com erro (i.e. ruído) desprezível
    - Exemplo: $y = \alpha + \beta x + \varepsilon$

# Regressão Linear Simples | Exemplo

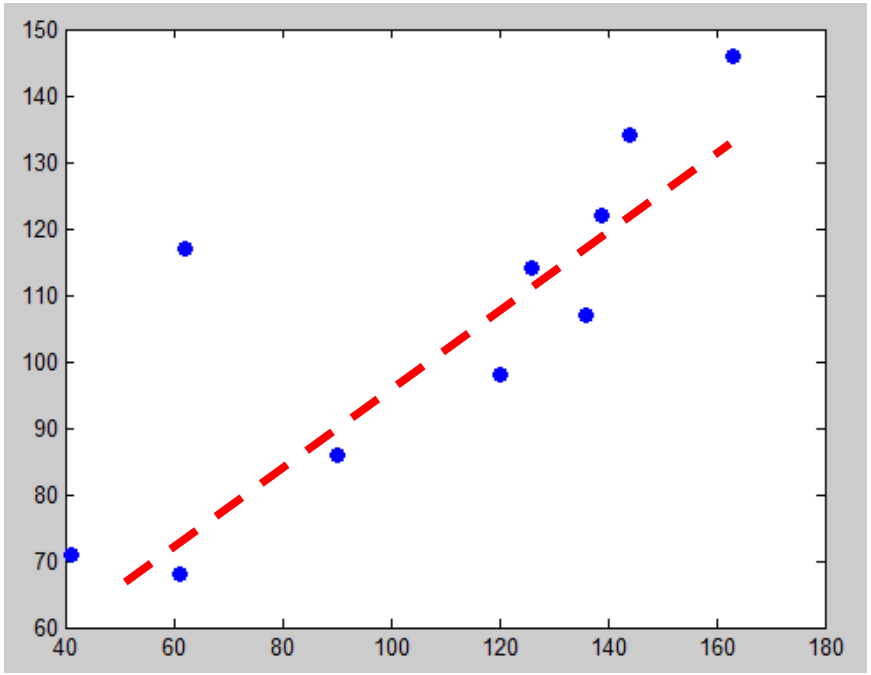
- Função
- y = 1,55 * x + 3,86



# Regressão Linear Simples

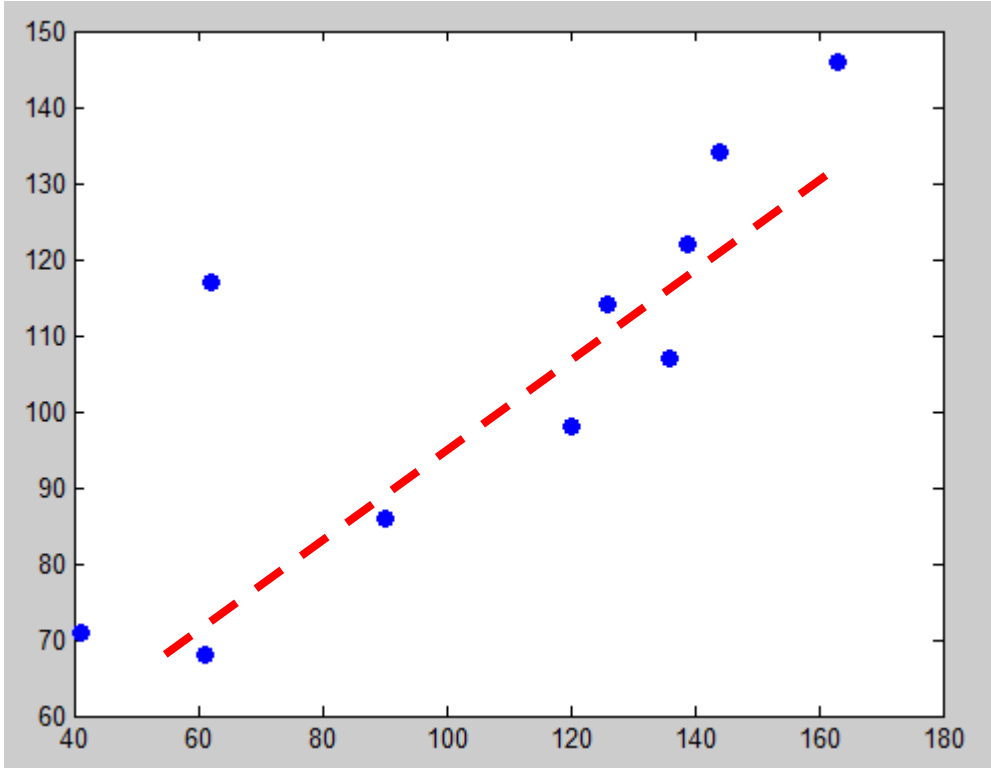
- A regressão implica no ajuste de uma reta que represente forma “adequada” a estrutura dos dados

| y | x |
|---|---|
| 122 | 139 |
| 114 | 126 |
| 86 | 90 |
| 134 | 144 |
| 146 | 163 |
| 107 | 136 |
| 68 | 61 |
| 117 | 62 |
| 71 | 41 |
| 98 | 120 |



# Regressão Linear Simples

- O que seria uma reta ajustada de forma “adequada”?
  - Reta com “menor distância possível” em relação aos valores observados
  - Para isso, devemos “**Minimizar a Soma dos Quadrados dos Resíduos**”



# Regressão Linear Simples

- Informações importantes
  - Na análise de regressão linear parte-se da suposição de que os erros (ou resíduos) têm distribuição normal
    - Média igual a zero e variância $\sigma_{\varepsilon}^{2}$
  - Os resíduos também podem ser escritos na forma $\varepsilon = y - \alpha - \beta x$

# Método dos Mínimos Quadrados

- Desenvolvimento
  - Proposto por Carl Friedrich Gauss em 1795
    - Utilizou o método no cálculo de órbitas de planetas e cometas a partir de medidas obtidas por telescópios
  - Adrien Marie Legendre publicou primeiro em 1806
    - Desenvolveu o mesmo método de forma independente

# Método dos Mínimos Quadrados

- O que é?
  - Técnica de otimização matemática
  - Procura o melhor ajuste para um conjunto de dados
    - ***(x(1),y(1)), (x(2),y(2)),..., (x(n),y(n))***
  - Ao mesmo tempo em que tenta minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados
    - $\sum_{i=1}^{n} \varepsilon_i^2$

# Método dos Mínimos Quadrados

- Objetivo
  - Procurar pelos parâmetros $\alpha$ e $\beta$ que minimizem a soma dos quadrados dos resíduos
    - $J(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \big(y(i) - \alpha - \beta x(i)\big)^2$
  - Isso equivale a fazer com que a soma dos quadrados dos resíduos entre os valores medidos (observações) e a reta de regressão seja mínima

# Método dos Mínimos Quadrados

- Equação de regressão
  - É calculada a partir das derivadas parciais da soma dos quadrados dos resíduos
  - Derivadas parciais com relação aos parâmetros $\alpha$ e $\beta$
    - $\frac{d}{d\alpha}J(\alpha,\beta) = -2\sum_{i=1}^{n}\big(y(i) - \alpha - \beta x(i)\big)^2$
    - $\frac{d}{d\beta}J(\alpha,\beta) = -2\sum_{i=1}^{n}\big(y(i) - \alpha - \beta x(i)\big)^2 x(i)$

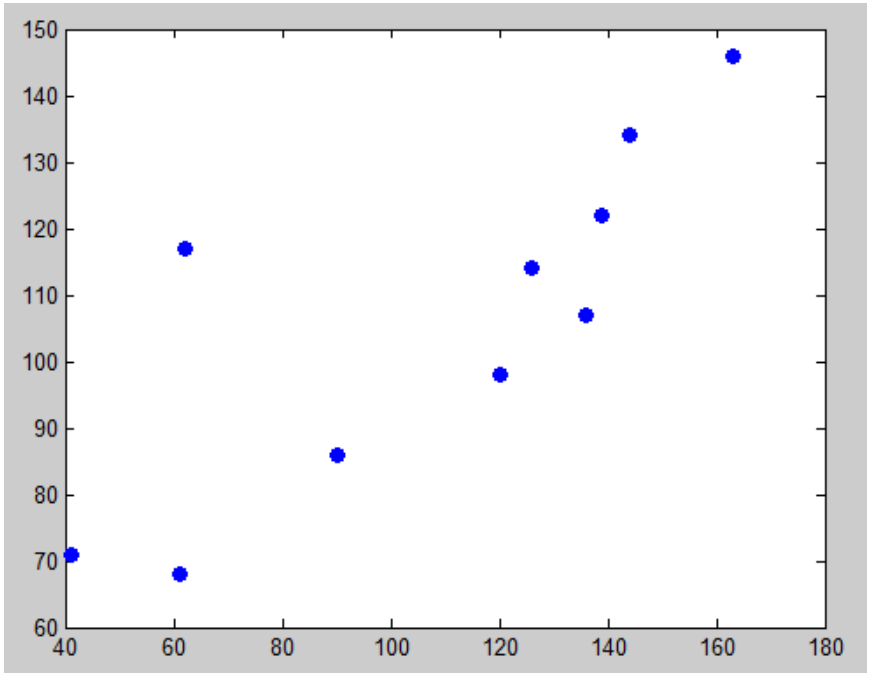
# Método dos Mínimos Quadrados

- Equação de regressão
  - Algumas deduções matemáticas e substituições depois e temos que
    - $\alpha = \bar{y} - \beta\bar{x}$
    - $\beta = \frac{\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})}{\sum_{i=1}^{n}(x(i)-\bar{x})^2}$
  - Onde $\bar{x}$ e $\bar{y}$ são as médias amostrais de ***x*** e ***y***, respectivamente

# Método dos Mínimos Quadrados | Exemplo

- Calcular a regressão para o seguinte conjunto de dados

| y | x |
|---|---|
| 122 | 139 |
| 114 | 126 |
| 86 | 90 |
| 134 | 144 |
| 146 | 163 |
| 107 | 136 |
| 68 | 61 |
| 117 | 62 |
| 71 | 41 |
| 98 | 120 |

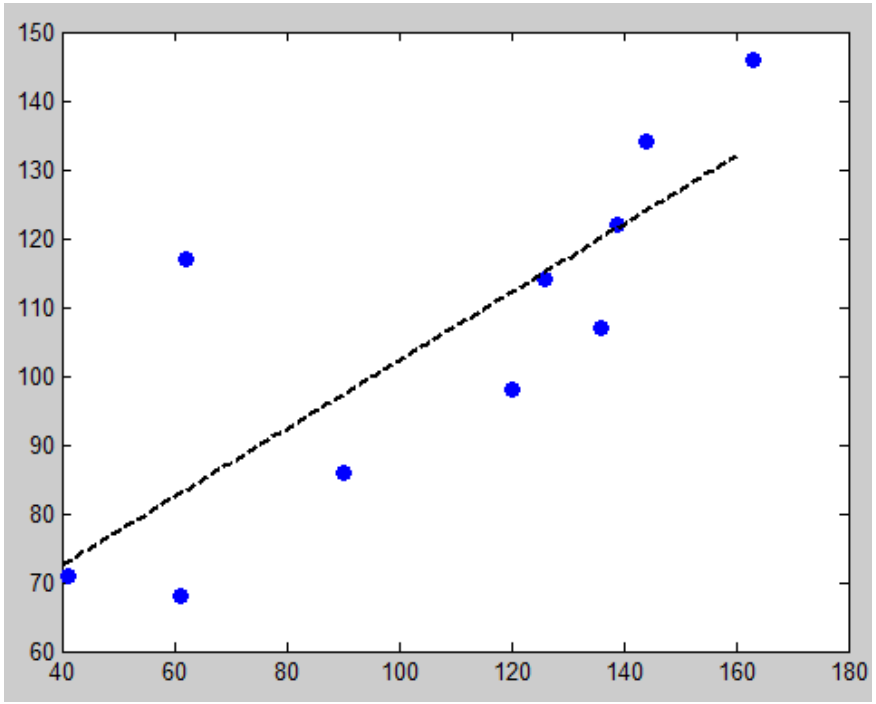


# Método dos Mínimos Quadrados | Exemplo

- Calcular a regressão para o seguinte conjunto de dados

| Média y | Média x |
|---|---|
| 106,3 | 108,2 |

- $\alpha = 52{,}69$
- $\beta = 0{,}4954$
- $y = 52{,}69 + 0{,}4954x$

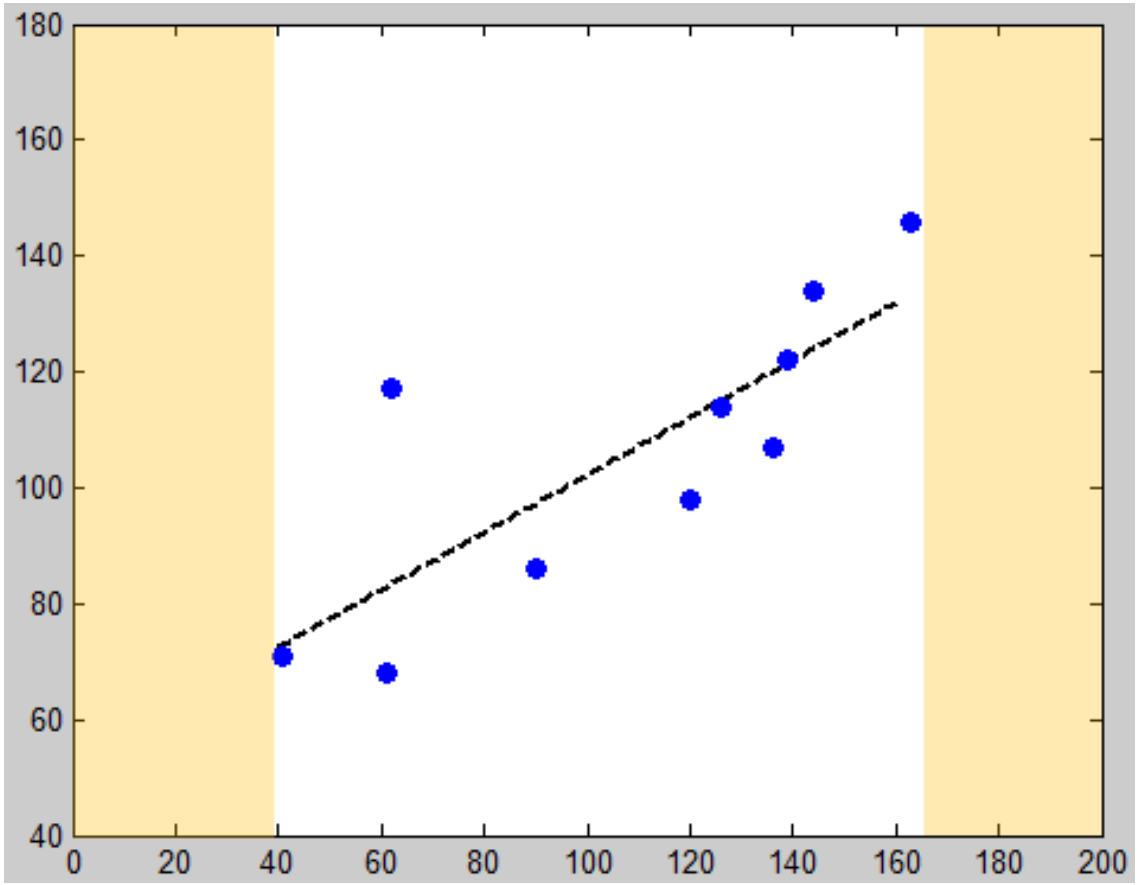


# Método dos Mínimos Quadrados

- Importante
  - Normalmente, a relação linear $y = \alpha + \beta x$ é considerada válida apenas para $x \in [x_{min}, x_{max}]$
    - Modelos de regressão linear não costumam ser válidos para fins de extrapolação, apenas de interpolação

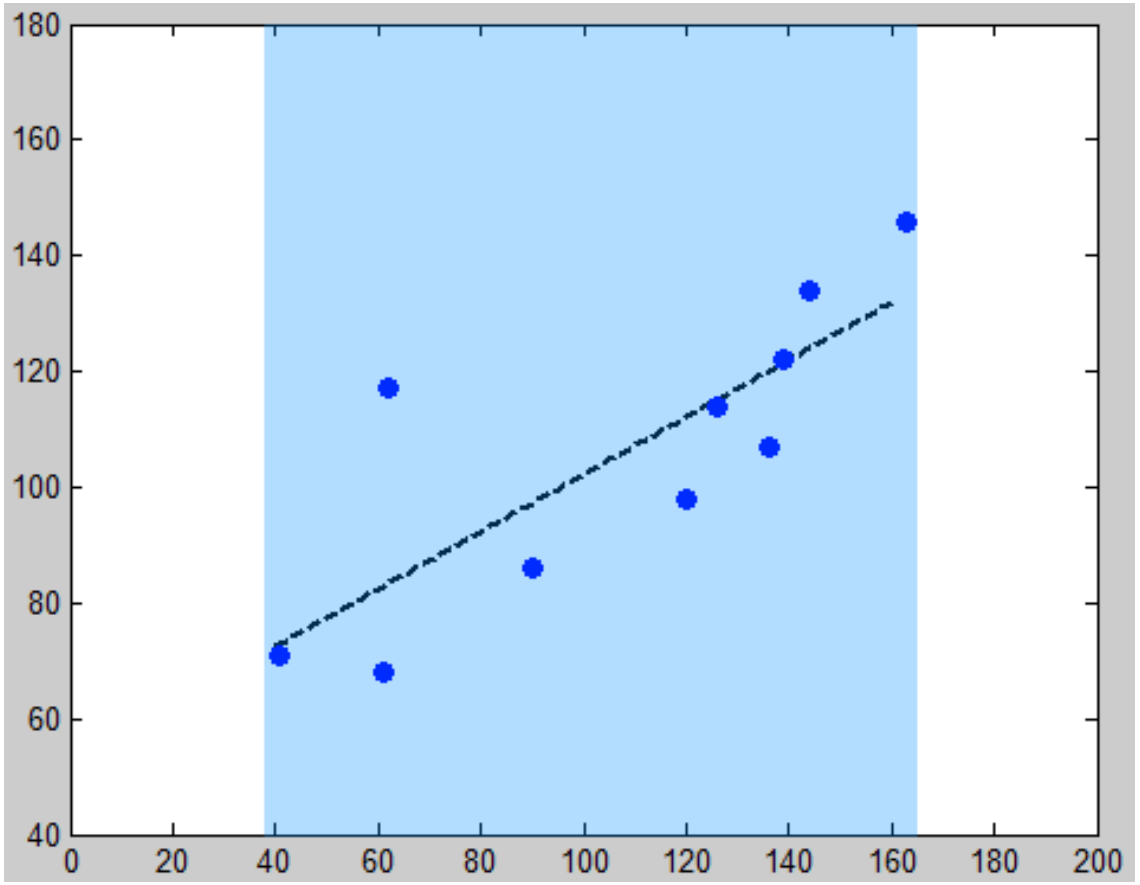
# Método dos Mínimos Quadrados

- Extrapolação
  - Calcular um valor de uma equação ou função, em um lugar fora da zona conhecida



# Método dos Mínimos Quadrados

- Interpolação
  - Calcular um valor de uma equação ou função, em um lugar da zona conhecida



# Análise de Resíduos

- Como podemos avaliar a qualidade do nosso modelo?
  - O modelo é adequado?
  - Os erros tem distribuição normal?
  - Os erros são independentes?
  - Os erros tem variância constante?
  - Por acaso existem valores discrepantes ?
    - Presença de outliers

# Análise de Resíduos

- Podemos fazer isso analisando os resíduos
  - Temos a disposição um conjunto de técnicas utilizadas para investigar o quão adequado um modelo de regressão está com base nos resíduos
  - O resíduo $e(i)$ é calculado como sendo a diferença entre nosso dado $y(i)$ e a sua estimativa $\hat{y}(i)$
    - $e(i) = y(i) - \hat{y}(i)$
    - $y(i) = \alpha + \beta x(i)$

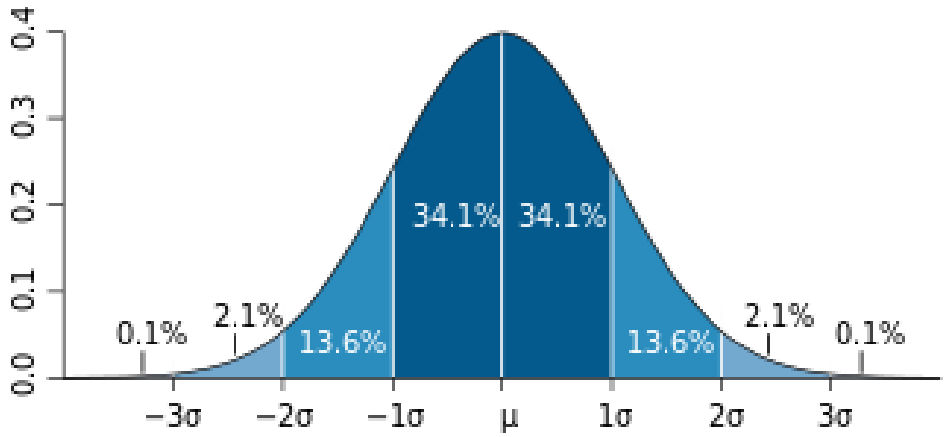
# Análise de Resíduos

- A análise dos resíduos permitem validar as suposições impostas pelo termo de erro do modelo e, portanto, adequado

- Suposições impostas
  - Média zero
  - Não correlacionados
  - Distribuição normal

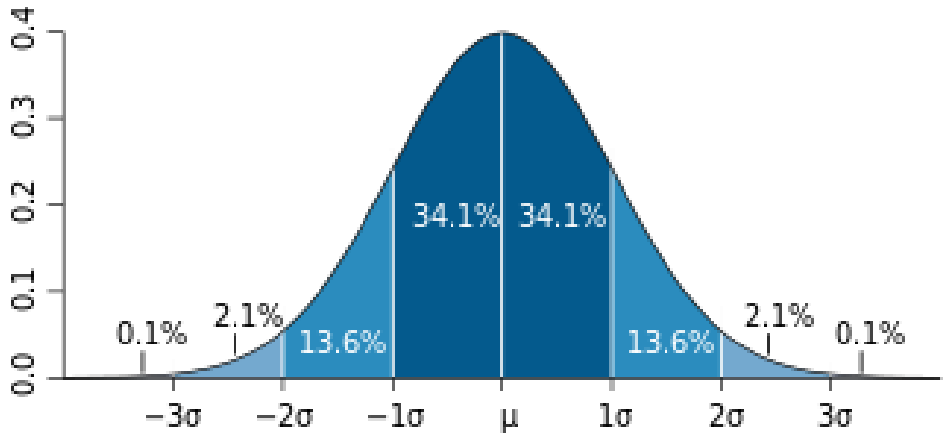
# Análise de Resíduos

- Presença de valores discrepantes ou outliers
  - Construir um histograma da frequência dos resíduos
    - Normalizar os resíduos: $d(i) = \frac{e(i)}{\widehat{\sigma_\epsilon}}$
  - O histograma dos resíduos deve ser semelhante a uma distribuição gaussiana



# Análise de Resíduos

- Presença de valores discrepantes ou outliers
  - Se os erros tiverem distribuição normal, então
    - Aproximadamente 95% dos resíduos normalizados devem cair dentro do intervalo (−2,+2)
    - Resíduos muito fora do intervalo (−2,+2) podem indicar a presença de um valor atípico em relação ao restante dos dados (outlier)

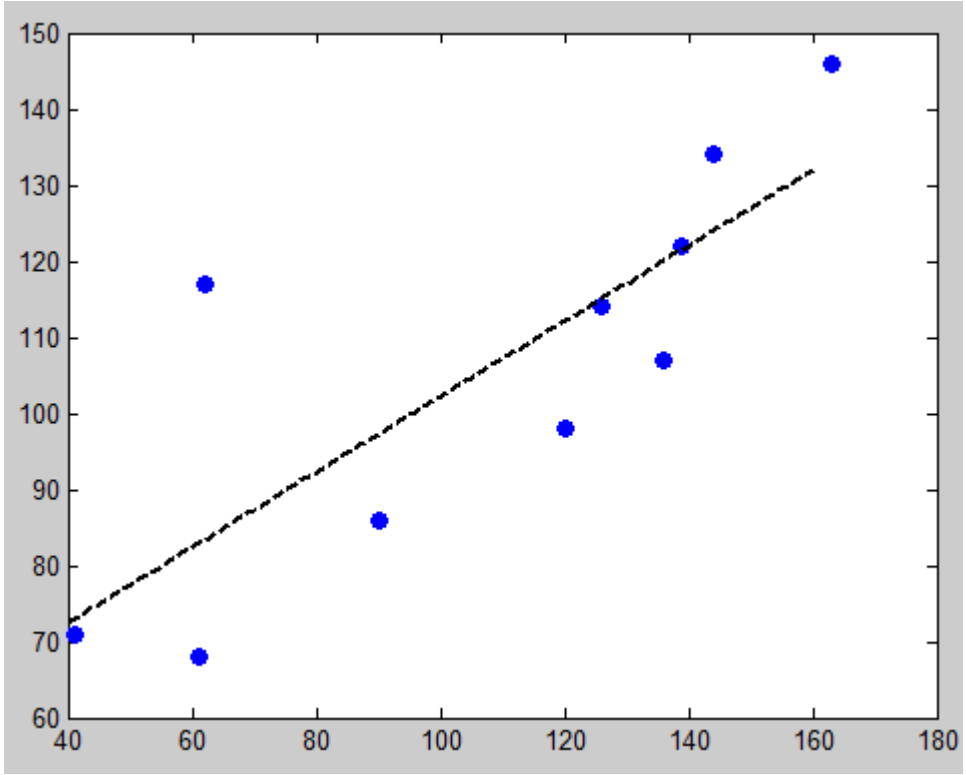


# Análise de Resíduos

- O que fazer com um outlier?
  - Alguns autores recomendam que eles sejam descartados
  - Outros autores acham que eles não devem ser descartados
    - Outliers fornecem informação importante sobre “falhas” e são de interesse para o experimentador

# Coeficiente de Determinação

- Observe a reta de regressão
  - Os pontos estão distribuídos acima e abaixo dela
  - O coeficiente de determinação, $R^2$, indica a quantidade de variabilidade dos dados que o modelo de regressão é capaz de explicar



# Coeficiente de Determinação

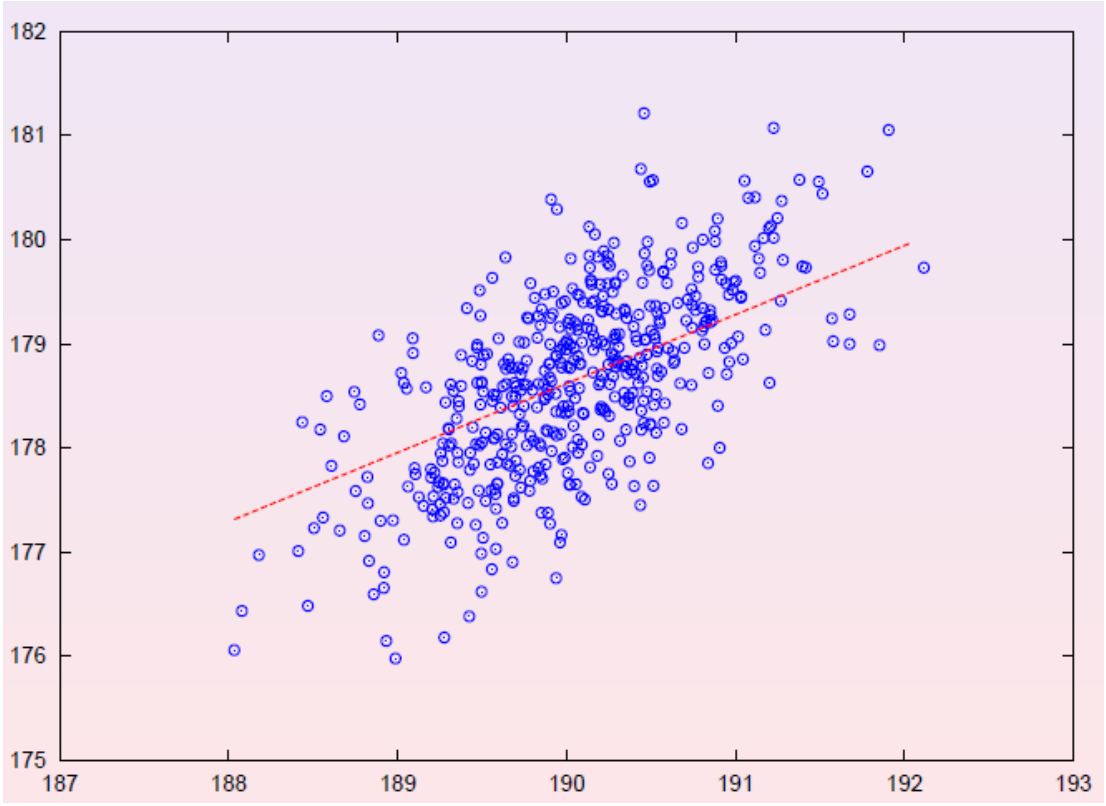
- Calculando $R^2$
  - O coeficiente de determinação é dado por

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y(i) - \hat{y}(i))^2}{\sum_{i=1}^{n}(y(i) - \bar{y}(i))^2}$$

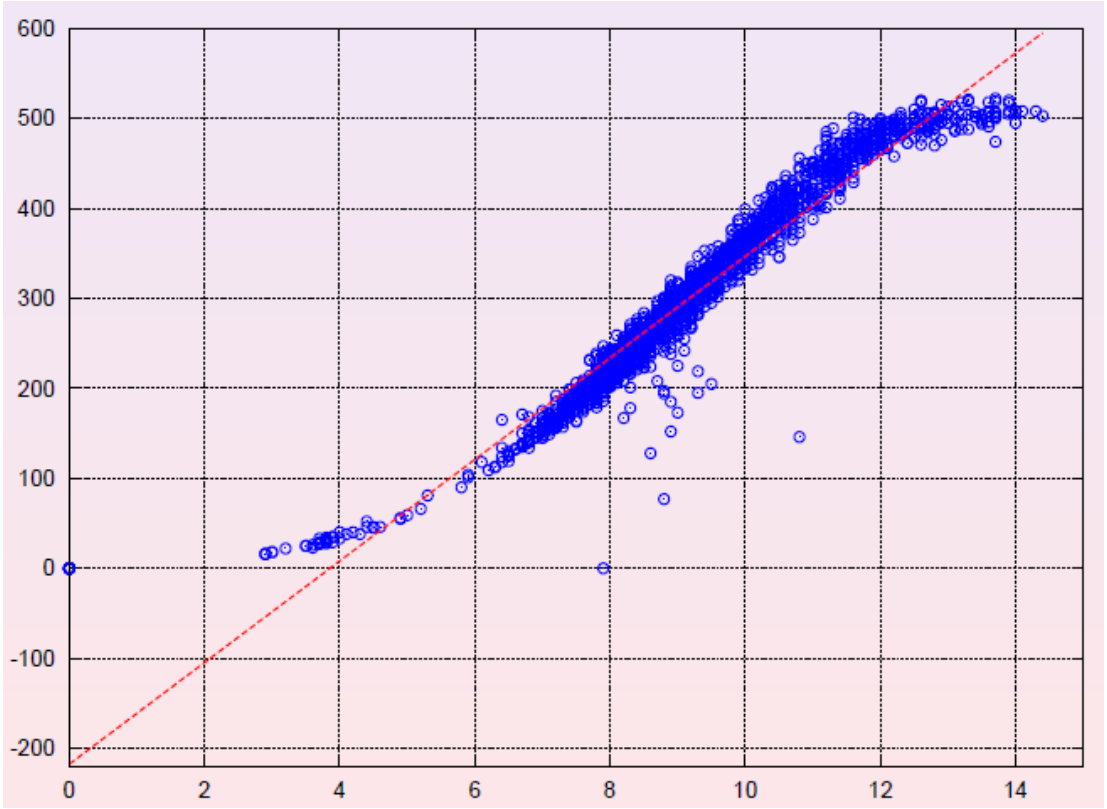
  - O valor resultante será $0 \leq R^2 \leq 1$
    - Quanto mais próximo o valor de $R^2$ está de 1, mais adequado é o modelo de regressão

# Coeficiente de Determinação

$R^2 = 0,44$

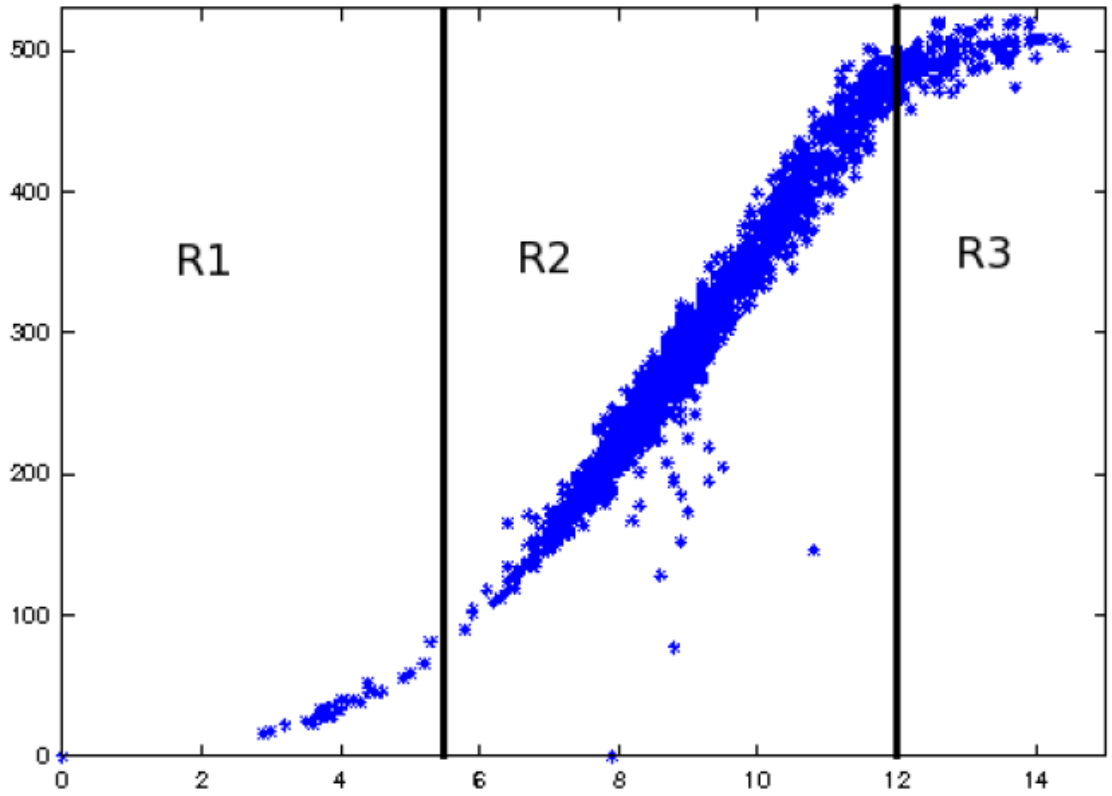



$R^2 = 0,93$



# Dados não lineares

- O que fazer  quando o modelo de regressão linear não é apropriado?
  - Solução 1:
    - Podemos dividir o domínio original dos dados em sub-domínios
    - Aplicar o modelo linear dentro de cada sub-domínios



# Dados não lineares

- O que fazer  quando o modelo de regressão linear não é apropriado?
  - Solução 2:
    - Podemos utilizar um modelo de regressão polinomial de ordem maior do que 1 ou não linear
    - Aplicar uma linearização dos dados e continuar usando a regressão linear

# Regressão não linear

- Definição
  - Forma de regressão em que os dados são modelados por uma função que é uma  combinação não linear de parâmetros
    - Pelo menos um dos seus parâmetros deve estar na forma não linear
  - Exemplos
    - Função exponencial: $y = \alpha e^{\beta x}$
    - Função logarítmica: $y = \alpha + \beta \log x$
    - Função de Potência: $y = \alpha x^{\beta}$

# Regressão não linear

- Por quê usar?
  - Muito importante na Biologia
  - Muitas aplicações biológicas são modeladas por meio de relações não lineares
    - Modelos de crescimento
    - Modelos de rendimento
    - Relações alométricas;

# Regressão não linear

- Como calcular a regressão?
  - Podemos tentar transformar uma relação não linear em linear (transformação linearizante)
    - Em seguida resolvemos o problemas como linear
  - Exemplo
    - Relação exponencial: $y = \alpha e^{\beta x}$
    - Modelada como: $y' = \alpha' + \beta x$
    - Onde $y' = \log y$ e $\alpha' = \log \alpha$

# Regressão não linear

- Como calcular a regressão?
  - Nem sempre é possível fazer essa transformação
    - Algumas relações não lineares não são linearizáveis
    - Estimar os parâmetros na relação linearizada não produz os mesmos resultados que estimar os parâmetros na relação não linear original

# Regressão não linear

- Como calcular a regressão?
  - Como na regressão linear, os dados são ajustados geralmente pelo método dos Mínimos Quadrados
    - Isso vale para relações linearizadas ou não
  - Ou podemos usar um método de aproximações sucessivas
    - Método de *Gauss-Newton*

# Regressão Linear Múltipla

- Idéia
  - A intuição nos diz que, geralmente, se pode melhorar uma predição se incluirmos novas variáveis independentes ao modelo (equação) de regressão
    - Uma reta é um polinômio de ordem 1
    - Usar de modelos polinomiais de ordem maior que 1

# Regressão Linear Múltipla

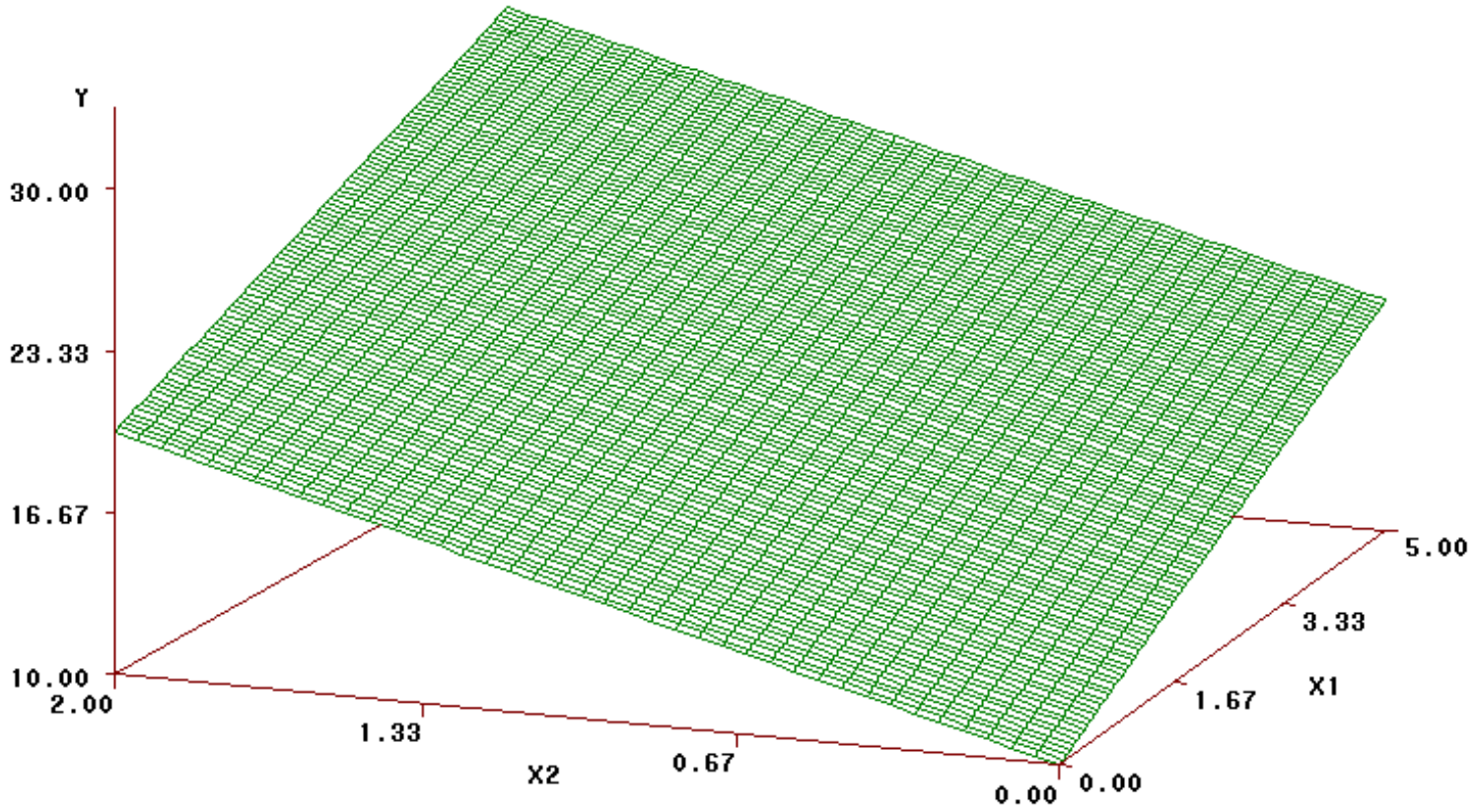
- Idéia
  - Antes de tudo devemos buscar o “equilíbrio” entre o número de parâmetros e a “capacidade preditiva” do modelo
  - Número excessivo de parâmetros
    - Sobreajustamento: modelo é muito específico
  - Número reduzido de parâmetros
    - Subajustamento: modelo pode ser pouco preditivo

# Regressão Linear Múltipla

- Idéia
  - A regressão múltipla funciona de forma parecida com a regressão simples
    - Basicamente, ela leva em consideração diversas variáveis de entrada $\boldsymbol{x_i}$, $\boldsymbol{i=1,...p}$, influenciando ao mesmo tempo uma única variável de saída, $\boldsymbol{y}$
  - Exemplo
    - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ ... + \beta_p x_p + \varepsilon$

# Regressão Linear Múltipla | Exemplo

- Função
- $y = 10 + 2x_1 + 5x_2$



# Regressão Linear Múltipla

- Informações importantes
  - A função de regressão na regressão múltipla é chamada de superfície de resposta
    - Ela descreve um hiperplano no espaço p-dimensional das variáveis de entrada $\boldsymbol{x_i}$
  - Os parâmetros $\beta_i, i = 0, \ldots, p$ são os coeficientes de regressão

# Regressão Linear Múltipla

- Vantagens
  - Permite representar modelos mais complexos e não apenas lineares
- Exemplo
  - Considere a seguinte equação de regressão com três variáveis de entrada
    - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

# Regressão Linear Múltipla

- Exemplo (continuação)
  - Se considerarmos
    - $x_1 = x$
    - $x_2 = x^2$
    - $x_3 = x^3$
  - Teremos escrito um modelo não linear (polinomial cúbico) em uma variável de entrada
    - $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

# Regressão Linear Múltipla

- Como calcular a superfície de regressão?
  - Usar o **método dos mínimos quadrados** como feito com a regressão linear simples
    - Ele pode ser usado para estimar os coeficientes de regressão $\beta_i, i = 0, \ldots, p$
  - Problema: elevado número de parâmetros
    - Temos ***n*** equações na forma $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots + \beta_p x_p + \varepsilon$, uma para cada observação dos dados

# Regressão Linear Múltipla

- Solução
  - Expressar as operações matemáticas utilizando notação matricial
    - $\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} \cdots & x_{p1} \\ 1 & x_{12} \cdots & x_{p2} \\ \cdots & \cdots \cdots & \cdots \\ 1 & x_{1p} \cdots & x_{pn} \end{pmatrix} * \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}$
    - $y = X\beta + e$

# Regressão Linear Múltipla

- Considerações importantes
  - Os erros (ou resíduos) têm distribuição normal
    - Média igual a zero e variância $\sigma_{\varepsilon}^{2}$
  - As observações não são correlacionadas
  - Temos ***n*** observações, sendo ***n>p***
    - Há mais equações do que incógnitas

# Regressão Linear Múltipla

- Método dos Mínimos Quadrados
  - A solução continua a mesma: procurar pelos parâmetros $\beta_i, i = 0, \ldots, p$ que minimizem a soma dos quadrados dos resíduos
    - $J(\beta) = \sum_{i=1}^{n} \varepsilon_i^2$
  - A equação acima pode ser reescrita como sendo
    - $J(\beta) = e'e$
  - Onde $e$ é o vetor de resíduos, e $e'$ é a sua transposta

# Regressão Linear Múltipla

- Método dos Mínimos Quadrados
  - Nosso objetivo é fazer com que a soma dos quadrados dos resíduos entre os valores medidos (observações) e a superfície de regressão seja mínima
  - Como $e = \ y - X\beta$, nosso objetivo se torna minimizar
    - $J(\beta) = e'e = (y - X\beta)'(y - X\beta)$

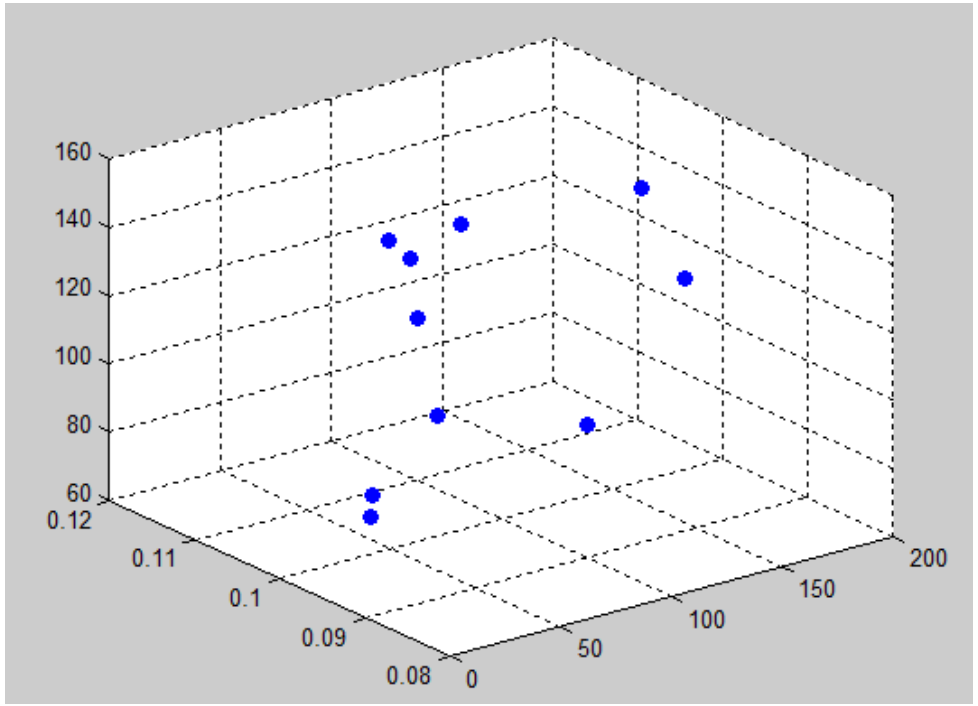
# Método dos Mínimos Quadrados

- Superfície de regressão
  - Algumas deduções matemáticas e substituições depois e temos que
    - $\beta = (X'X)^{-1}X'y$
  - Onde $A^{-1}$ representa a matriz inversa da matriz $A$

# Método dos Mínimos Quadrados | Exemplo

- Calcular a regressão para o seguinte conjunto de dados

| y | $x_1$ | $x_2$ |
|---|---|---|
| 122 | 139 | 0,115 |
| 114 | 126 | 0,12 |
| 86 | 90 | 0,105 |
| 134 | 144 | 0,09 |
| 146 | 163 | 0,1 |
| 107 | 136 | 0,12 |
| 68 | 61 | 0,105 |
| 117 | 62 | 0,08 |
| 71 | 41 | 0,1 |
| 98 | 120 | 0,115 |



# Método dos Mínimos Quadrados | Exemplo

- Calcular a regressão para o seguinte conjunto de dados

$$\beta = \left( \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 139 & 126 & 90 & \cdots \\ 0{,}115 & 0{,}12 & 0{,}105 & \cdots \end{pmatrix} \times \begin{pmatrix} 1 & 139 & 0{,}115 \\ 1 & 126 & 0{,}12 \\ 1 & 90 & 0{,}105 \\ \cdots & \cdots & \cdots \end{pmatrix} \right)^{-1} \times \begin{pmatrix} 1 & 1 & 1 & \cdots \\ 139 & 126 & 90 & \cdots \\ 0{,}115 & 0{,}12 & 0{,}105 & \cdots \end{pmatrix} \times \begin{pmatrix} 122 \\ 114 \\ 86 \\ \cdots \end{pmatrix}$$

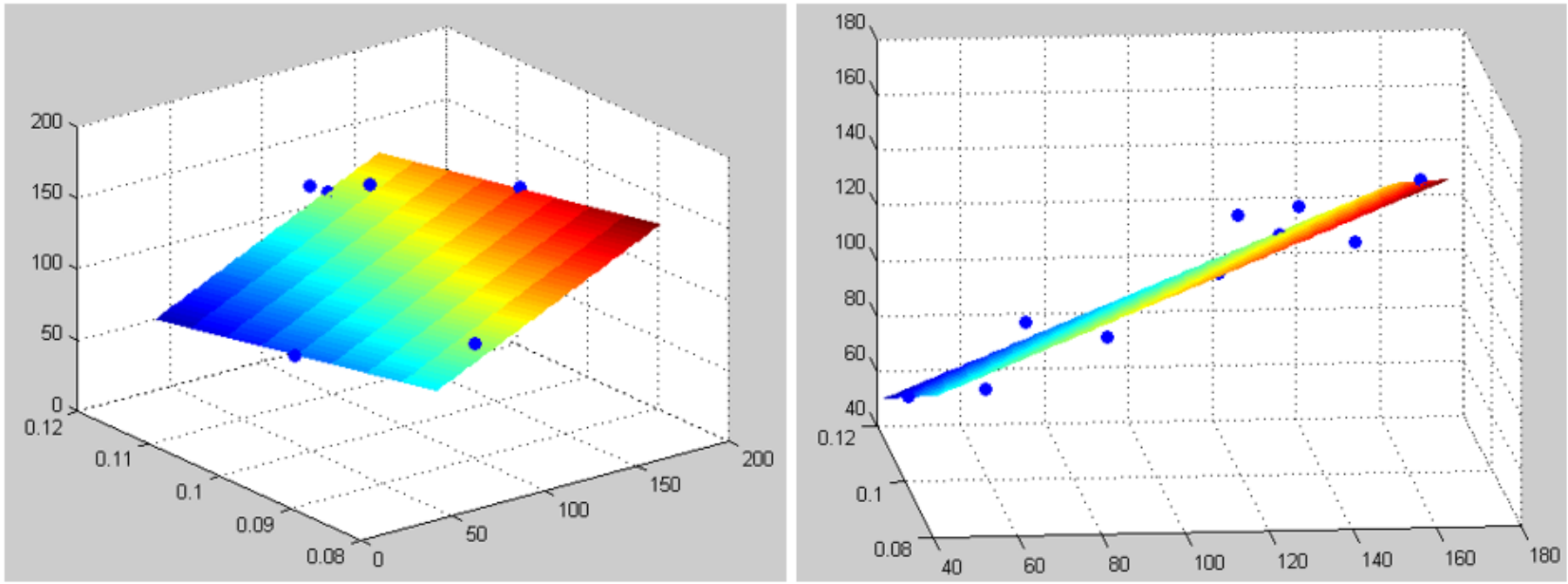
| y | $x_1$ | $x_2$ |
|---|---|---|
| 122 | 139 | 0,115 |
| 114 | 126 | 0,12 |
| 86 | 90 | 0,105 |
| 134 | 144 | 0,09 |
| 146 | 163 | 0,1 |
| 107 | 136 | 0,12 |
| 68 | 61 | 0,105 |
| 117 | 62 | 0,08 |
| 71 | 41 | 0,1 |
| 98 | 120 | 0,115 |

# Método dos Mínimos Quadrados | Exemplo

- Calcular a regressão para o seguinte conjunto de dados
- Solução do sistema
  - $\beta = \begin{pmatrix} 148{,}52 \\ 0{,}6136 \\ -1034{,}41 \end{pmatrix}$
  - $y = 148{,}52 + 0{,}6136x_1 - 1034{,}41x_2$

# Método dos Mínimos Quadrados | Exemplo

- Calcular a regressão para o seguinte conjunto de dados



# Regressão Linear Múltipla

- Problemas
  - Nem sempre é possível calcular a inversa da matriz $(X'X)^{-1}$
    - Seu determinante muitas vezes é zero ou quase igual a zero
    - Isto geralmente ocorre quando as variáveis de entrada são intercorrelacionadas
    - Se a intercorrelação é grande existe ***multicolinearidade***: as linhas da matriz $X'X$ não são linearmente independentes

# Multicolinearidade

- Como minimizar esse efeito?
  - Aplicar a regularização de Tikhonov
    - A regressão passa a ser chamada de **regressão de cumeeira** (*ridge regression*).
  - A equação usada para calcular os parâmetros $\beta$
    - $\beta = (X'X)^{-1}X'y$
  - É reescrita da seguinte forma
    - $\hat{\beta} = (X'X + \lambda \boldsymbol{I})^{-1}X'y$

# Multicolinearidade

- Como minimizar esse efeito?
  - Basicamente, com a regularização de Tikhonov, nós somamos uma constante a diagonal principal da matriz de modo a tentar torná-la inversivel.
    - $\hat{\beta} = (X'X + \lambda \boldsymbol{I})^{-1} X'y$
  - Onde
    - $0 \leq \lambda \ll 1$ é uma constante de valor pequeno
    - $\boldsymbol{I}$ é uma matriz identidade de ordem ***(p+1)***

# Coeficiente de Determinação

- O coeficiente de determinação também pode ser obtido para uma regressão múltipla

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y(i) - \hat{y}(i))^2}{\sum_{i=1}^{n}(y(i) - \bar{y}(i))^2}$$

- Problema
  - Nesse caso, um valor alto para $R^2$ não significa que o modelo seja bom

# Coeficiente de Determinação

- Por que $R^2$ alto não significa “bom”?
  - Acrescentar uma variável ao modelo sempre aumentará o valor de $R^2$ , mesmo que a variável adicional não seja significante (informativa)
  - O que fazer então?
    - Podemos calcular o ***coeficiente de determinação ajustado***

# Coeficiente de Determinação

- Coeficiente de determinação ajustado

$$R^2_{aj} = 1 - \frac{\sum_{i=1}^{n}(y(i) - \hat{y}(i))^2 \Big/ (n-k)}{\sum_{i=1}^{n}(y(i) - \bar{y}(i))^2 \Big/ (n-1)}$$

- Onde $k = p + 1$
  - Desse modo, o valor do coeficiente de determinação irá crescer apenas se a adição de um novo termo reduzir significantemente a média quadrática dos erros

# Agradecimentos


- Agradeço ao professor Guilherme de Alencar Barreto da Universidade Federal do Ceará (UFC) pelo material disponibilizado