

Regressões

Danilo Augusto Cleto Souto
danilosouto@gmail.com

Apresentação

- Danilo Augusto Cleto Souto
 - Sistemas de Informação / Integração de Sistemas / Modelos Preditivos / Inteligência Artificial e Analytics
 - Celepar
 - 9 Anos
 - Linguagens
 - **Java**, C, C++, PHP , **Python** e **R**
 - GSUS
 - Laboratório
 - Automação laboratorial
 - Telefonia
 - Core Telefonia IP
 - Tarifação
 - SMS Mais de 600.000 mês
 - Aplicações de IA para o Governo do Paraná

Quem Somos?

- Nome?
- Formação?
- Onde Trabalha?
 - Utilizam modelos de predição? Quais?
 - Conhece a Linguagem R?
- O que faz?
 - Como gostaria que fosse feito?
 - Resultados?

Agenda

- Conceito de Regressão
- Para que serve?
- Tipos de regressão
- Básico de Linguagem R
- Regressão Linear Simples
- Erro
- Coeficiente de Correlação
- Regressão Linear Múltipla
- Regressão Logística
- Modelos Lineares Generalizados

Agenda

- Regressão Linear
- Básico Linguagem R
- Prática de R
- Teórica Regressão Linear Múltipla
- Modelos Lineares Generalizados
- Trabalho Final





Introdução

- Usada para criar **modelos** que demonstrem um fenômeno em observação;
- Um dos métodos estatísticos mais utilizados para investigar **relação** entre variáveis;

Ok, Mas para que serve?

- Para fazer previsões (Predições)
 - Previsão de populações
 - Demanda por serviços públicos.
- Para simular os efeitos entre duas variáveis:
 - Simular a evolução do IDH sobre investimentos em segurança ou educação.
 - Simular a produção de uma proteína animal em relação à ração adicionada.
- Para resumir dados:
 - **Milhões de valores** podem ser descritos somente por **uma reta**

Variáveis

- Sempre teremos pelo menos duas variáveis
 - X_1, \dots, X_n : Variável independente ou explicativa
 - Y : Variável dependente ou resposta
- Relação pode ser Direta ou Indireta
 - Direta (+): a variação de Y corresponde diretamente com a variação de X .  
 - Inversa (-): o valor de Y responde inversamente à variação de X  

Dependente

Independente

- Peso Corporal Animal ← Quantidade de Ração
- Produção Leiteira ← Quantidade de Proteína
- Produção Agrícola ← Fertilizante
- Eficácia (insetos mortos) ← Concentração de Inseticida
- Carga no sistema ← Quantidade Acessos

Regressão

- Prever os valores de uma variável dependente em função de uma variável independente.

- O que?

$$Y = f(X)$$

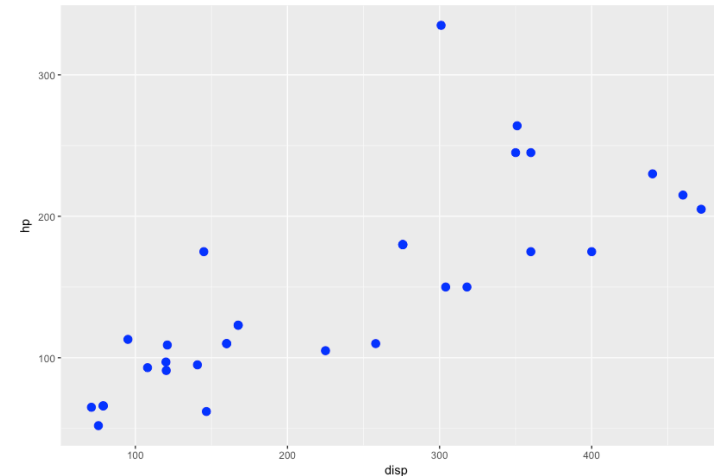
- Assim criamos uma fórmula que nos diz o quanto a **variável resposta (dependente)** irá ser afetada pelas variações da variável **explicativa (independente)**
- **Simplificar** nossos dados

Tipos de Modelos de Regressão

- Modelo de Regressão Simples (1 Variável)
 - Linear
 - Não Linear
- Modelo de Regressão Múltipla (mais de 1 variável)
 - Linear
 - Não Linear

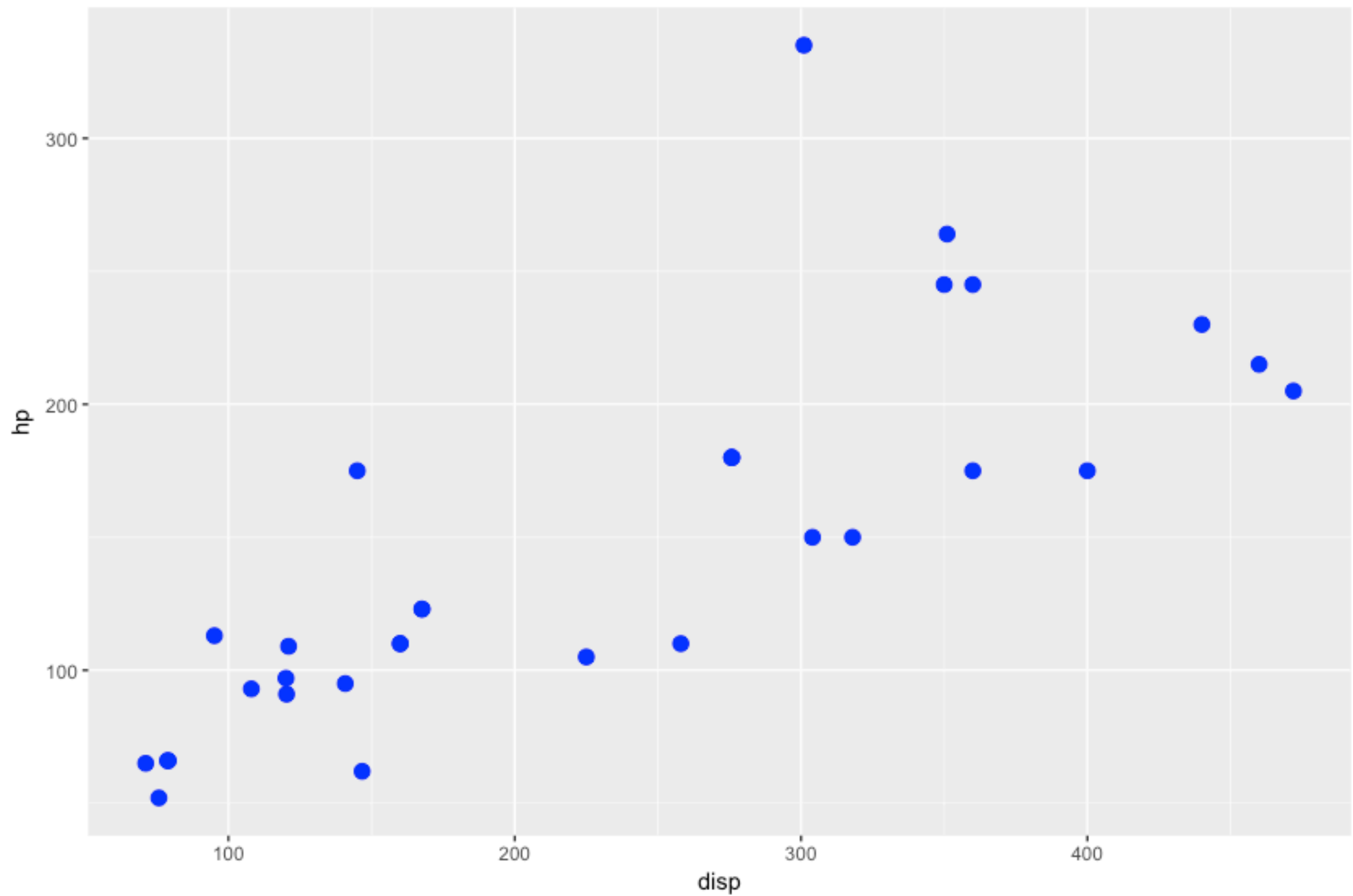
Mas como saber se existe relação

1. Gráfico de Dispersão:

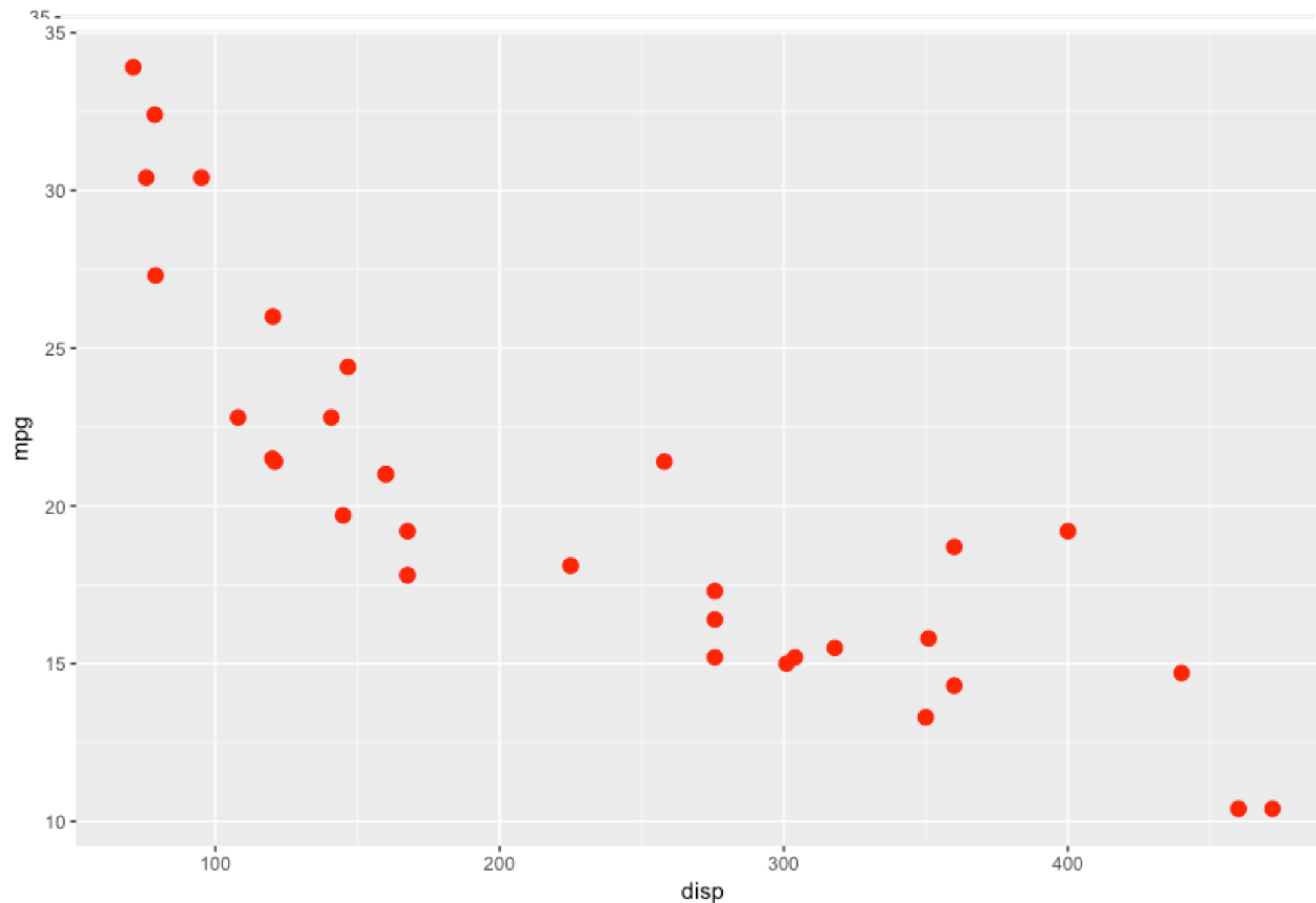


- Deve ser feito antes da análise dos dados. Com ele pode ser fácil perceber a existência de correlação entre variáveis.
- É construído de utilizando as variáveis em pares (X,Y)
- Constitui de uma nuvem de pontos onde é possível ter uma visão se existe alguma relação entre as variáveis.

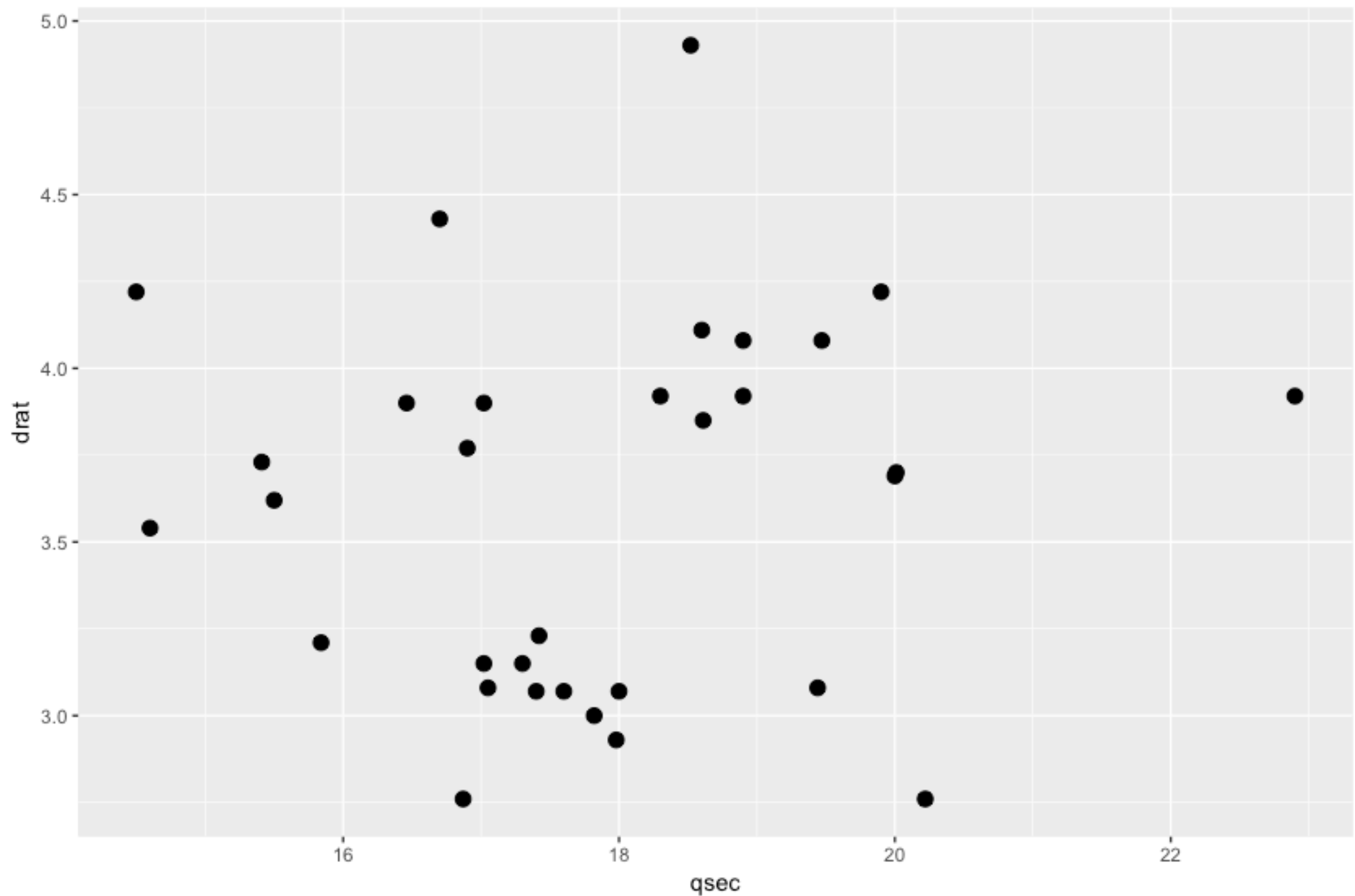
Correlação Positiva



Correlação Negativa



Ausência de Correlação



Linear Simple

- Modela a relação entre X e Y
- O valor Estimado (Y) como uma função do Valor Observado (X)
- Os valores nem sempre são iguais, por isso há sempre um **erro** associado

Linear Simples

$$Y = f(X)$$

$$Y = \alpha_0 + \alpha_1 X + e$$

Erro = e

Regressão Linear Simples

- Análise de regressão é definir os parâmetros α_0 e α_1 da função $f(x)$

$$Y = \alpha_0 + \alpha_1 X$$

$$y_i = \alpha_0 + \alpha_1 x_i$$

- Onde **alfa 0** é chamado ponto de **intersecção em Y** e **alfa 1** define a **inclinação** da reta.
- São os valores para correlacionarem as variáveis X e Y.
- É chamada **Reta de Regressão**;

Erro ou Desvio

- Como há sempre uma diferença entre os dados previstos e os valores observados, associamos sempre um erro ou desvio a equação onde

$$e_i = y_i - \hat{y}_i$$

- Mas o que é esse erro?

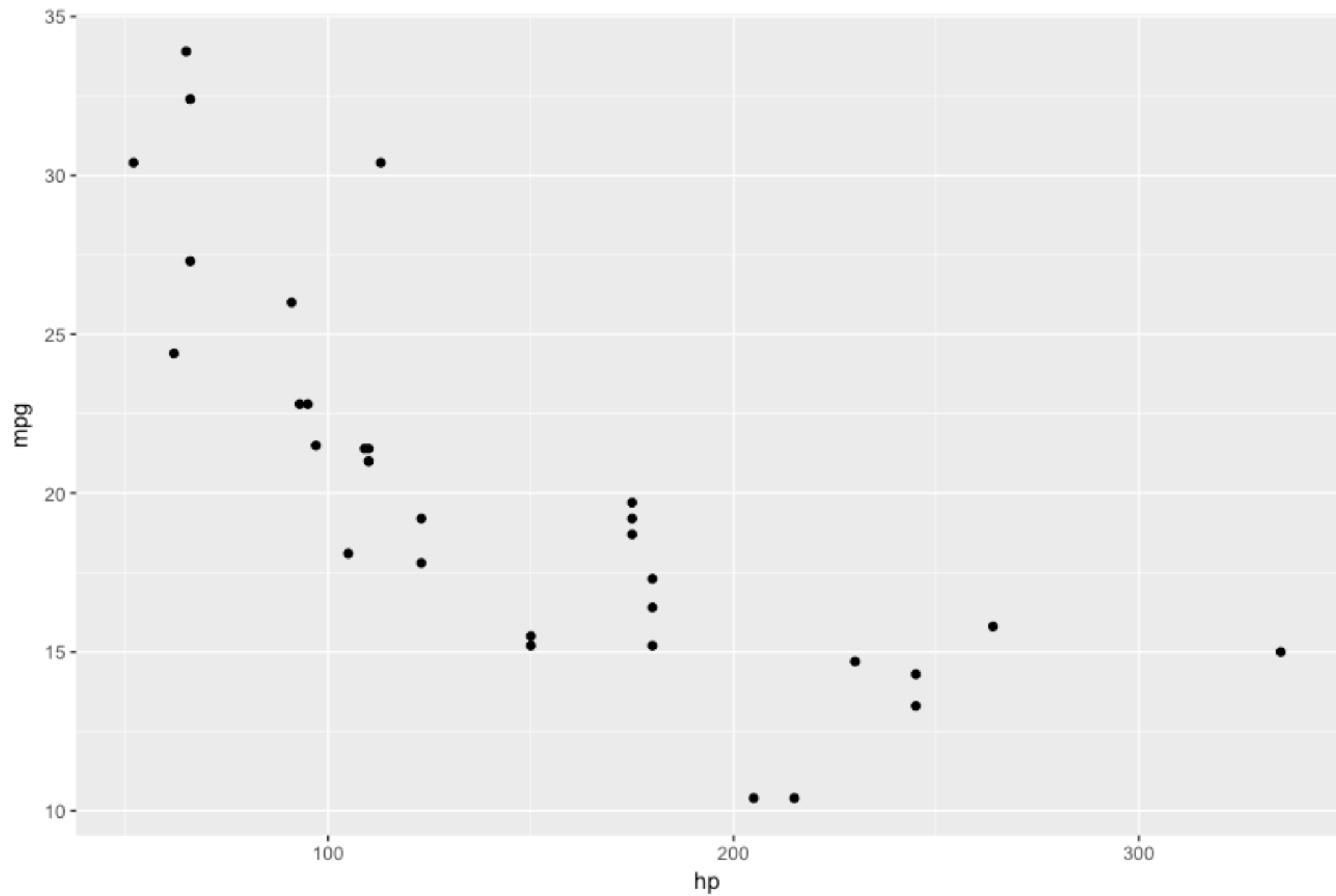
$$\hat{y}_i = y_i - e_i$$

- Indica que as variações de Y **não** são completamente **explicadas** por X;
- Podem existir **outras variáveis** que Y depende;
- A amostra podem pertencer a **outro grupo**;

Modelo Linear

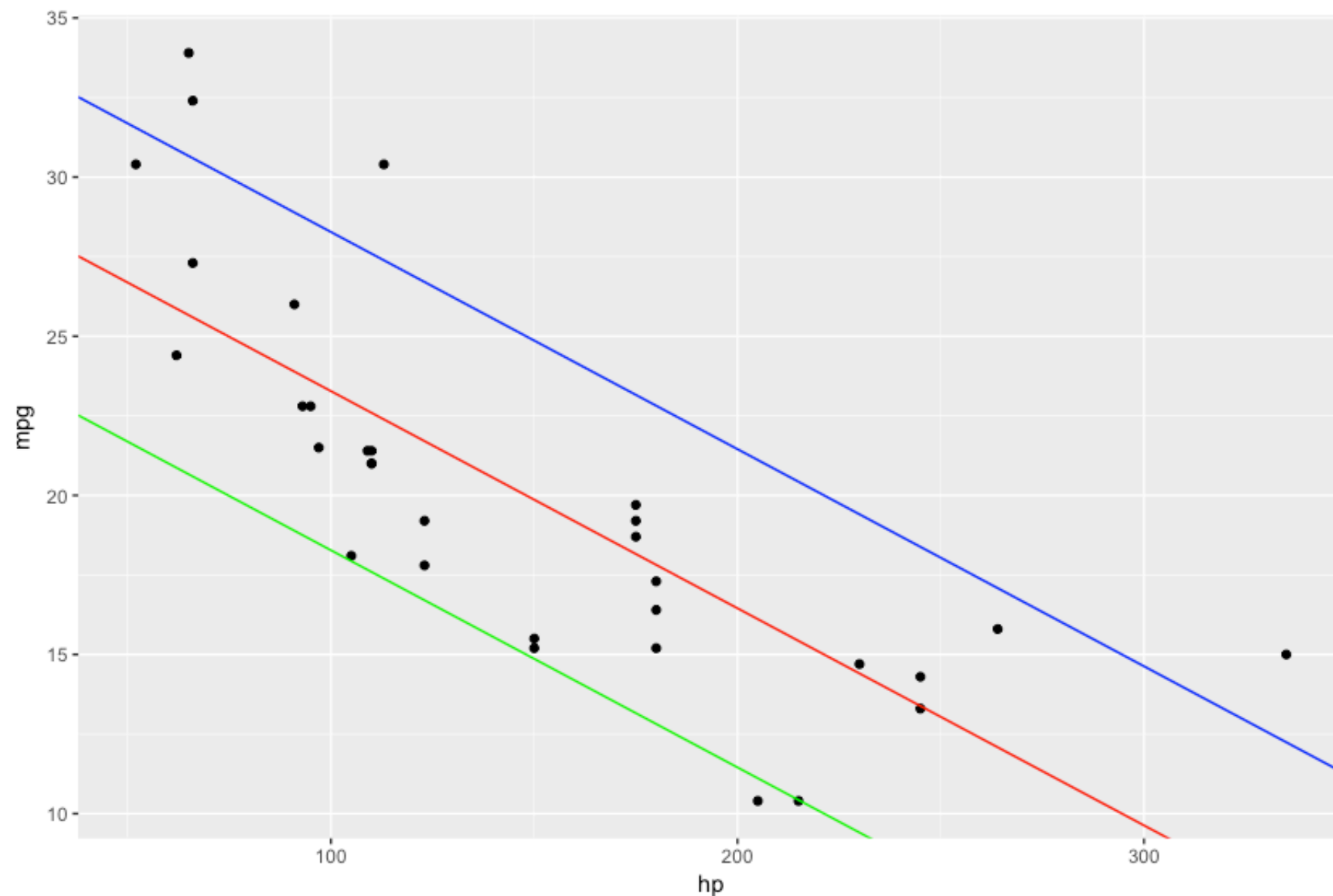
	Consumo (mpg)	Potência (hp)
Mazda RX4	21.0	110
Mazda RX4 Wag	21.0	110
Datsun 710	22.8	93
Hornet 4 Drive	21.4	110
Hornet Sportabout	18.7	175
Valiant	18.1	105
Duster 360	14.3	245
Merc 240D	24.4	62
Merc 230	22.8	95
Merc 280	19.2	123
Merc 280C	17.8	123
Merc 450SE	16.4	180
Merc 450SL	17.3	180
Merc 450SLC	15.2	180

Dados



Erros existem, mas o que fazer?

Nosso objetivo é ter uma reta que represente os nossos dados mas com a menor distância entre os dados estimados e os observados.



Minimizar o erro

- Encontrar uma equação onde a diferença entre os valores observados e os estimados seja o menor possível.

↓
$$e_{tot} = \sum_{i=1}^n (y_i - \hat{y}_i) = (y_1 - \hat{y}_1) + (y_2 - \hat{y}_2) + \dots + (y_n - \hat{y}_n)$$

$$y_i = \alpha_0 + \alpha_1 x_i + e_i \quad \downarrow$$

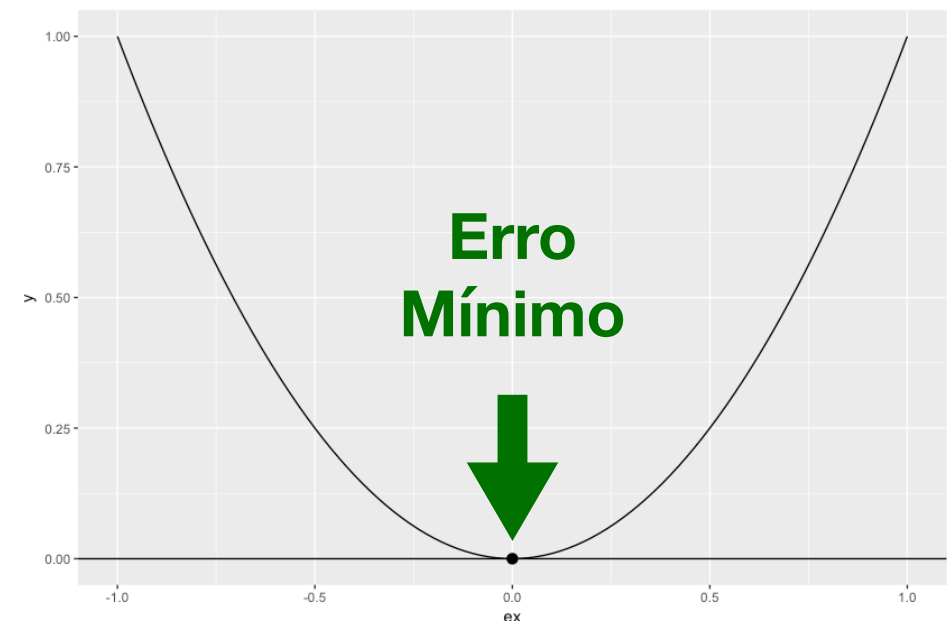
- Ou seja, definir os valores de **Alfa 0** e **Alfa 1** onde a soma dos **erros** esteja no **menor valor**.

Como minimizar o erro?

- Aplicar uma função para minimizar

$$e = Y - \alpha_0 - \alpha_1 X$$

↓ $(e)^2 = (Y - \alpha_0 - \alpha_1 X)^2$



- Os valores de **alfa 0** e **alfa1** para qual o erro seja o menor possível. Para encontrar o ponto do erro mínimo fazemos derivadas parciais em relação **alfa 0** e **alfa 1**

$$\frac{\partial e^2(\alpha_0, \alpha_1)}{\partial \alpha_0} = 0$$

$$\frac{\partial e^2(\alpha_0, \alpha_1)}{\partial \alpha_1} = 0$$

Linear Simple

$$\hat{Y} = \alpha_0 + \alpha_1 X$$

$$\alpha_1 = \frac{(X - \bar{X})(Y - \bar{Y})}{(X - \bar{X})^2}$$

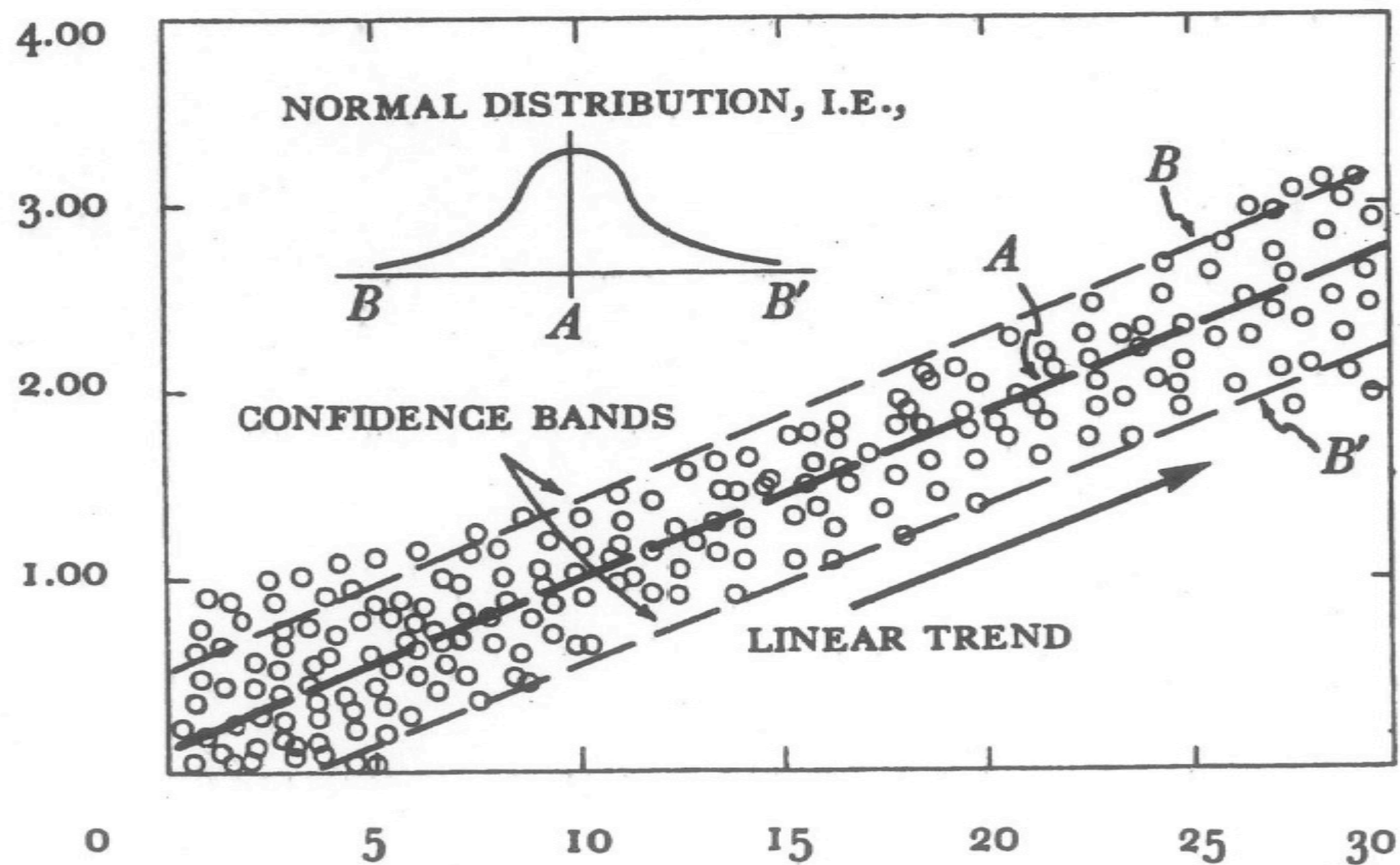
$$\alpha_0 = \bar{Y} - \alpha_1 \bar{X}$$

- Mínimos quadrados $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \dots + (y_n - \hat{y}_n)^2$
- A reta **simplifica** os dados e **minimiza** as diferenças (erro) é chamada de **Curva de Regressão**
- Assim a Curva de Regressão é uma tendência uma forma de simplificar os dados em apenas uma **equação simples** e capaz de estimar **valores não coletados**.

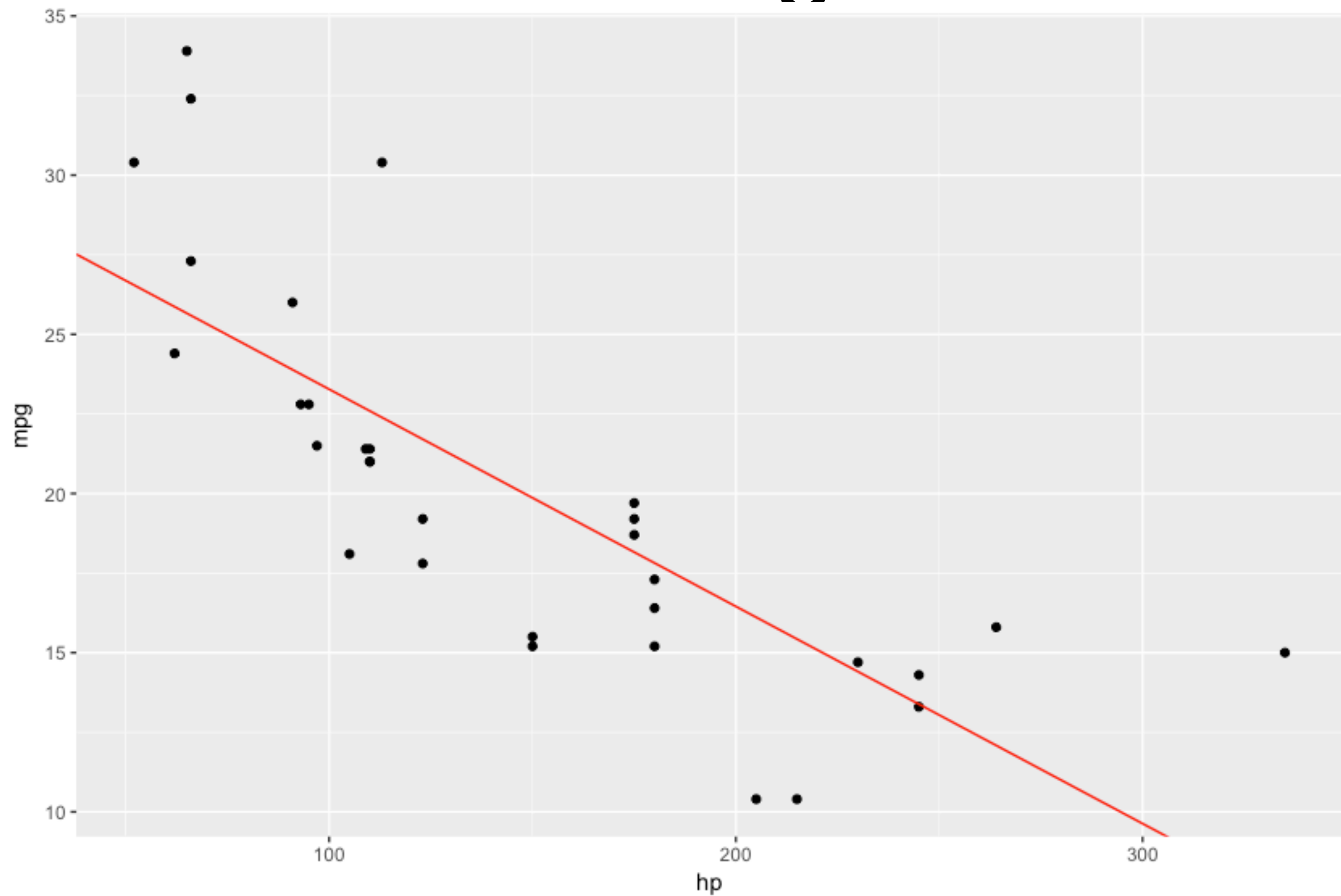
Erro padrão

- Mede o desvio entre Y e \hat{Y} na mesma unidade de Y .
- Pode ser medido como desvio padrão dos resíduos.
- Os valores observados (Y) seguem uma distribuição normal em torno do valor estimado (\hat{Y})

Regressão Linear



Curva de Regressão

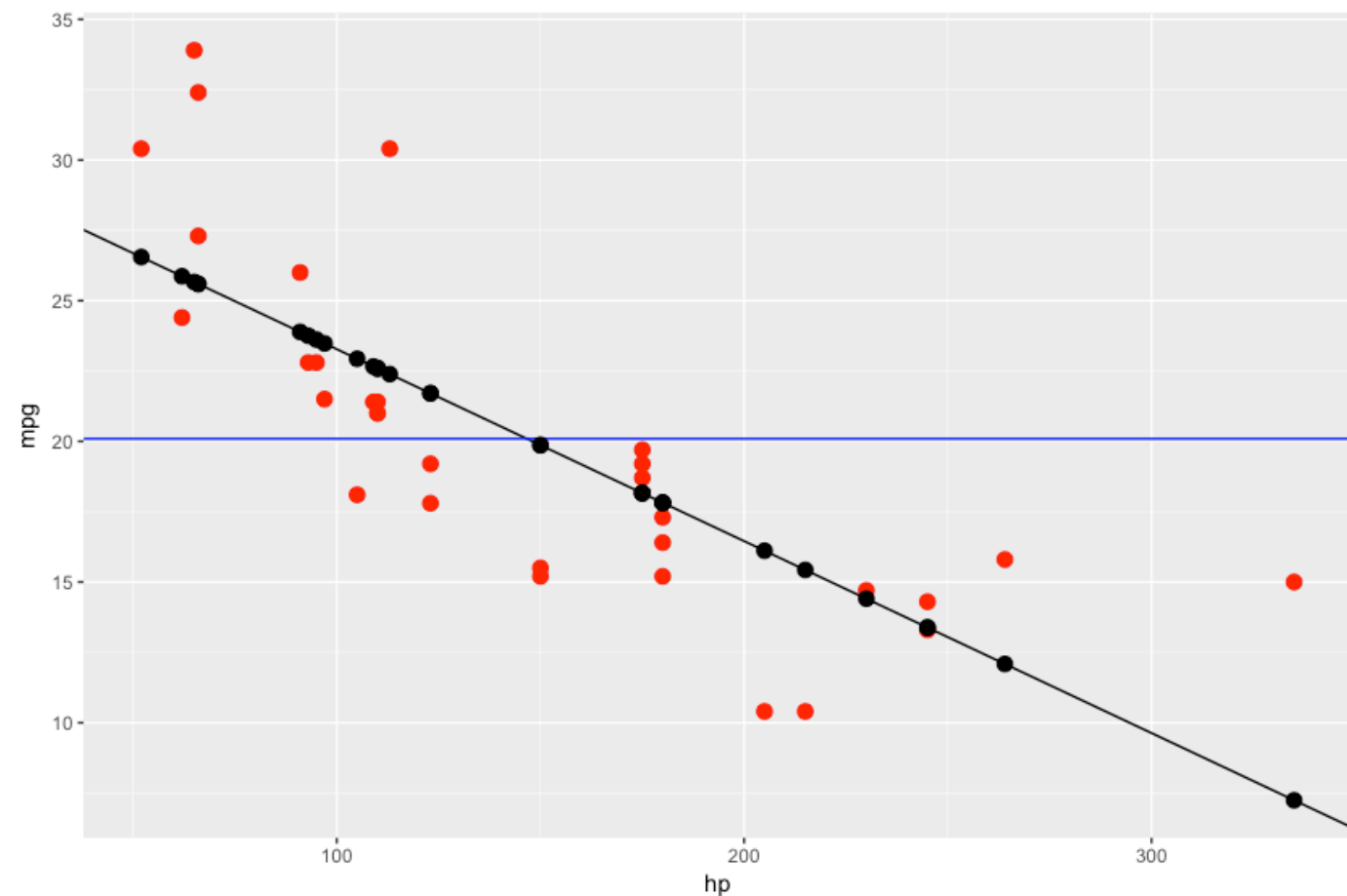


- Os valores estão próximos à reta de regressão. Para cima e para baixo;

- Valor Coletado

- Valor Estimado

- Valor Médio



Exercício

	X	Y
1	1	5.943180
2	2	7.733549
3	3	10.739441
4	4	13.255065

$$\hat{Y} = \alpha_0 + \alpha_1 X$$

$$\alpha_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

$$\alpha_0 = \bar{Y} - \alpha_1 \bar{X}$$

Exercício

	X	Y	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	1	5.719762	-1.5	-4.135058	2.25	6.2025874
2	2	8.384911	-0.5	-1.469909	0.25	0.7349546
3	3	11.779354	0.5	1.924534	0.25	0.9622669
4	4	13.535254	1.5	3.680434	2.25	5.5206506
SOMA	10	39.42	0	0	5	13.42

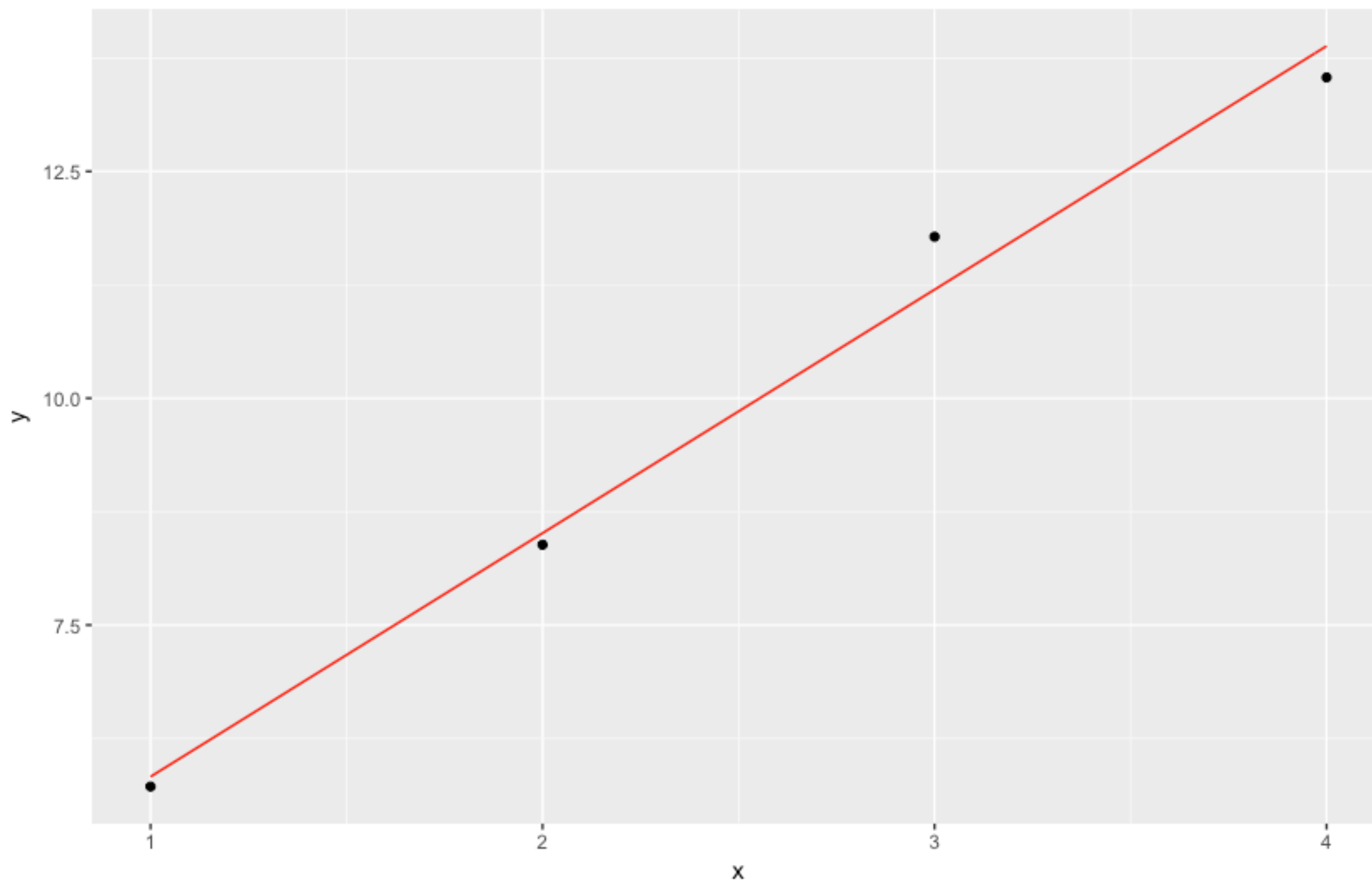
$$\bar{Y} = 9.85$$

$$\bar{X} = 2.5$$

$$\alpha_0 = \bar{Y} - \alpha_1 \bar{X}$$

$$\alpha_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

$$y = 3.14 + 2.68 x$$



Qualidade

- Mas como determinar a qualidade da minha aproximação?
 - Erro padrão
 - Coeficiente de Determinação
 - Coeficiente de Determinação Ajustado

Erro (Diferença)

- Variação total

- Coletado - Valor Médio

$$SQ_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$$

- Variação não explicada $Y - Y'$

- Coletado - Estimado

$$SQ_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Variação explicada

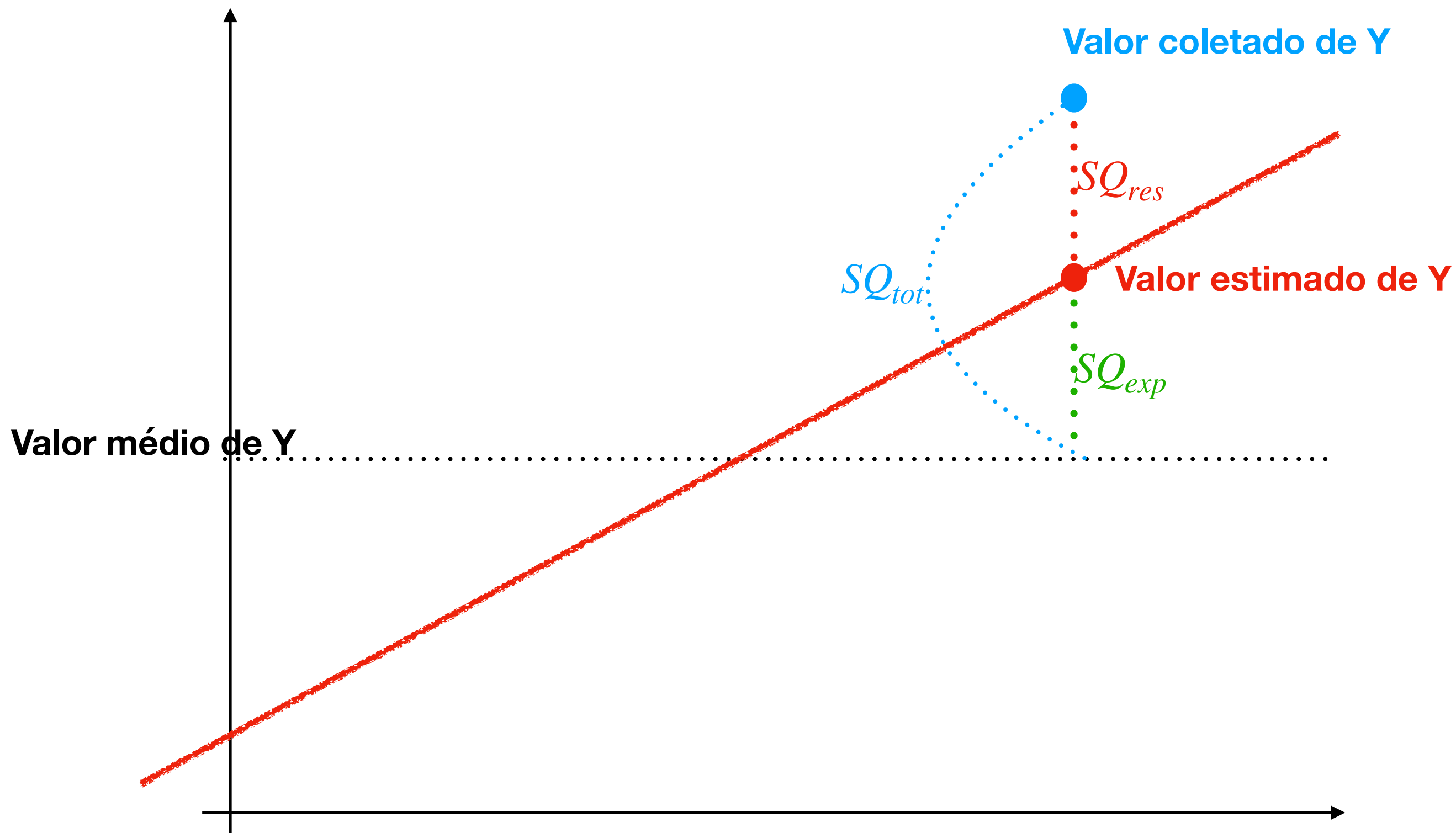
- Estimado - Valor Médio

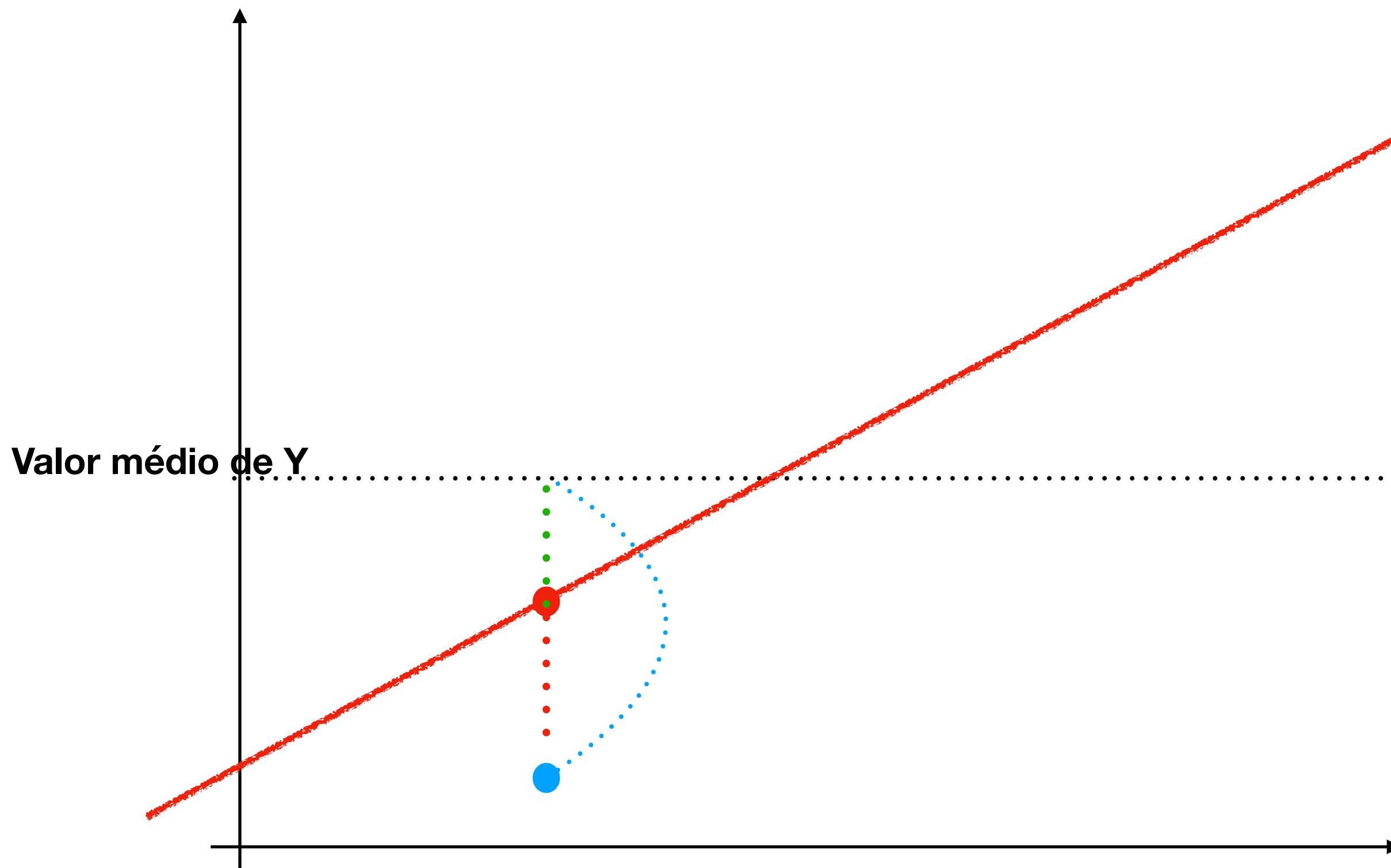
$$SQ_{exp} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

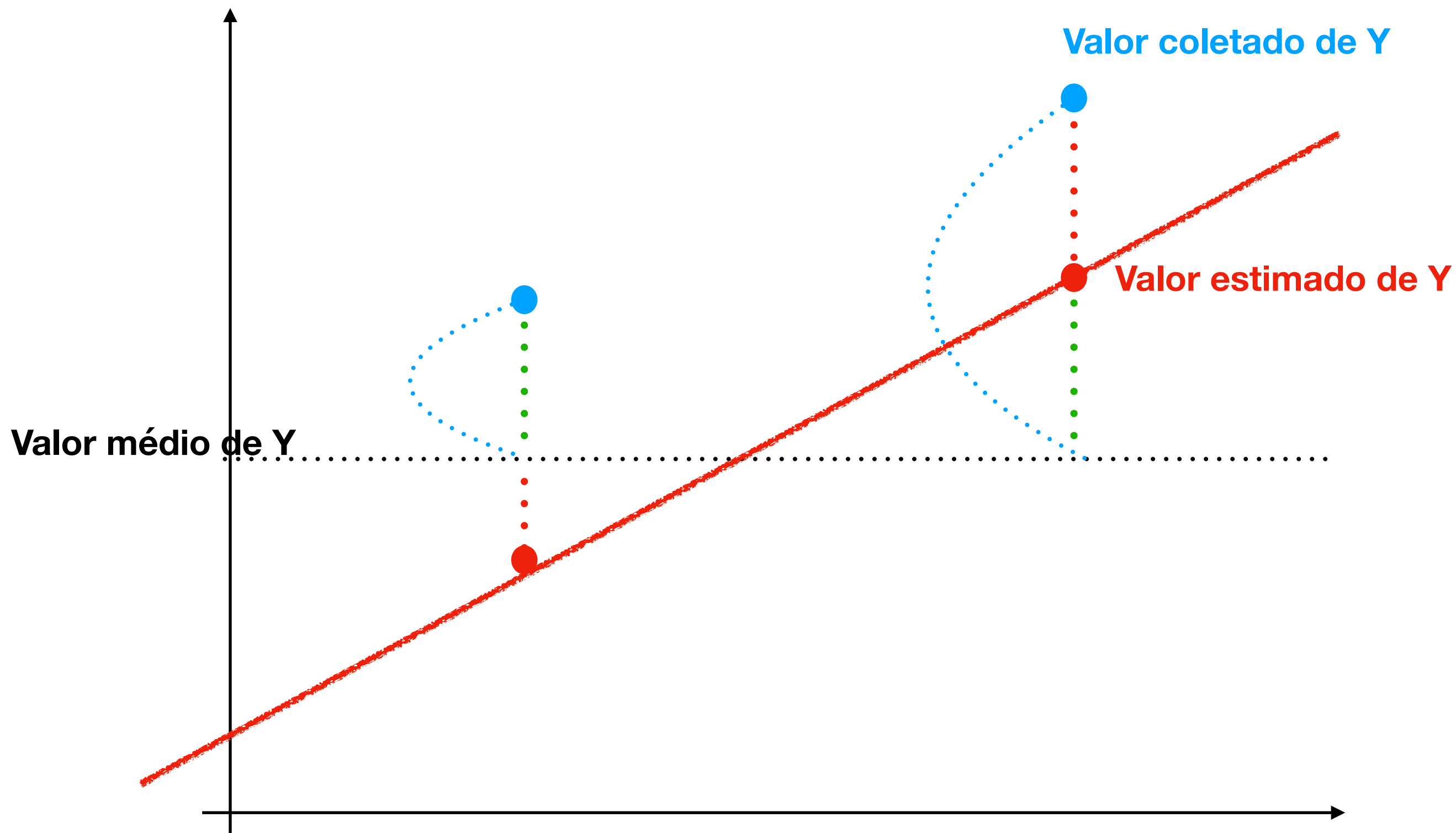
Variação Total = Variação Não Explicada + Variação Explicada

$$SQ_{tot} = SQ_{res} + SQ_{exp}$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$







Coeficiente de Determinação

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

$$r^2 = \frac{SQ_{exp}}{SQ_{tot}}$$

- É a razão (**proporção**) da variável dependente (**Y**) que é **explicado** pela variável independente (**X**).
- Em outras palavras, é o quanto a variável a **Curva de Regressão explica a variável Y**.

- Coeficiente de Correlação Simples (r)
- O coeficiente de determinação é sempre positivo e mede o quanto a curva se adequa aos dados
- O coeficiente de correlação assume valores positivos e negativos indicando correlação Direta (+) ou Inversa (-)

$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$cor(x, y) = \frac{Cov(x, y)}{\sigma_y \sigma_x}$$

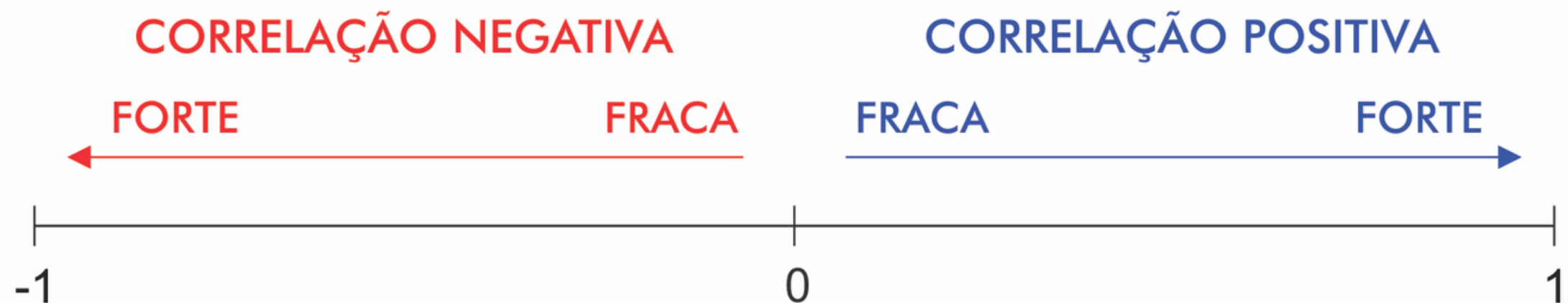
Coeficiente de Correlação

- Mede a força de **correlação** entre a as variáveis.
- Como entender o coeficiente de correlação.
- Pode assumir valores entre $-1 < r < 1$
- Próximos a **-1** entende-se por uma **correlação negativa**
- Próximos 0 ausência de correlação.
- Próximos a **1** uma forte **correlação positiva**.

Correlação

- $r = -1$: Correlação perfeita negativa
- $-1 < r < 0$: Correlação Negativa
- $r = 0$: Ausência de correlação
- $0 < r < 1$: Correlação Positiva
- $r = 1$: Correlação perfeita positiva
- $0,2 < r < 0,4$: *Correlação fraca*
- $0,4 < r < 0,7$: *Correlação moderada*
- $0,7 < r < 1$: *Correlação forte*

Correlação



Correlação de Pearson

- Avalia a relação **linear** entre duas variáveis **contínuas**.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Correlação de Spearman

- Diferentemente de Pearson pode ser usado mesmo quando as variáveis parecem não se comportar e quando as variáveis **não são contínuas**.
- Pode ser usado mesmo quando as variáveis não acompanham uma taxa constante.
- Uma simplificação quando temos duas variáveis inteiras e distintas podemos usar a seguinte aproximação:

$$r = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

- Onde **d** representa a diferença do **posto** de x em relação ao **posto** de y no ponto. E n é a quantidade de valores.

Spearman

- Ordenar os valores de acordo com x
- Adicionar Postos índices de 1 até n; (dx_i)
- Ordenar os os postos de acordo com a ordem de y (dy_i)
- Calcular diferença ($dx_i - dy_i = d_i$)
- Somar os quadrados de d_i .
- Aplicar a fórmula simplificada:

$$r = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Regressão Linear Múltipla

- Quando há relação com **mais de uma variável independente**
- Na verdade não será apenas uma reta unidimensional, mas sim, uma regressão **n-dimensional**.
 - Plano se forem 2 variáveis
 - Hiperplano se forem mais de 2 variáveis
- O que foi visto no modelo linear também é válido para o modelo multivariado.
- A regressão Linear Simples é uma regressão Linear Múltipla de 1 dimensão

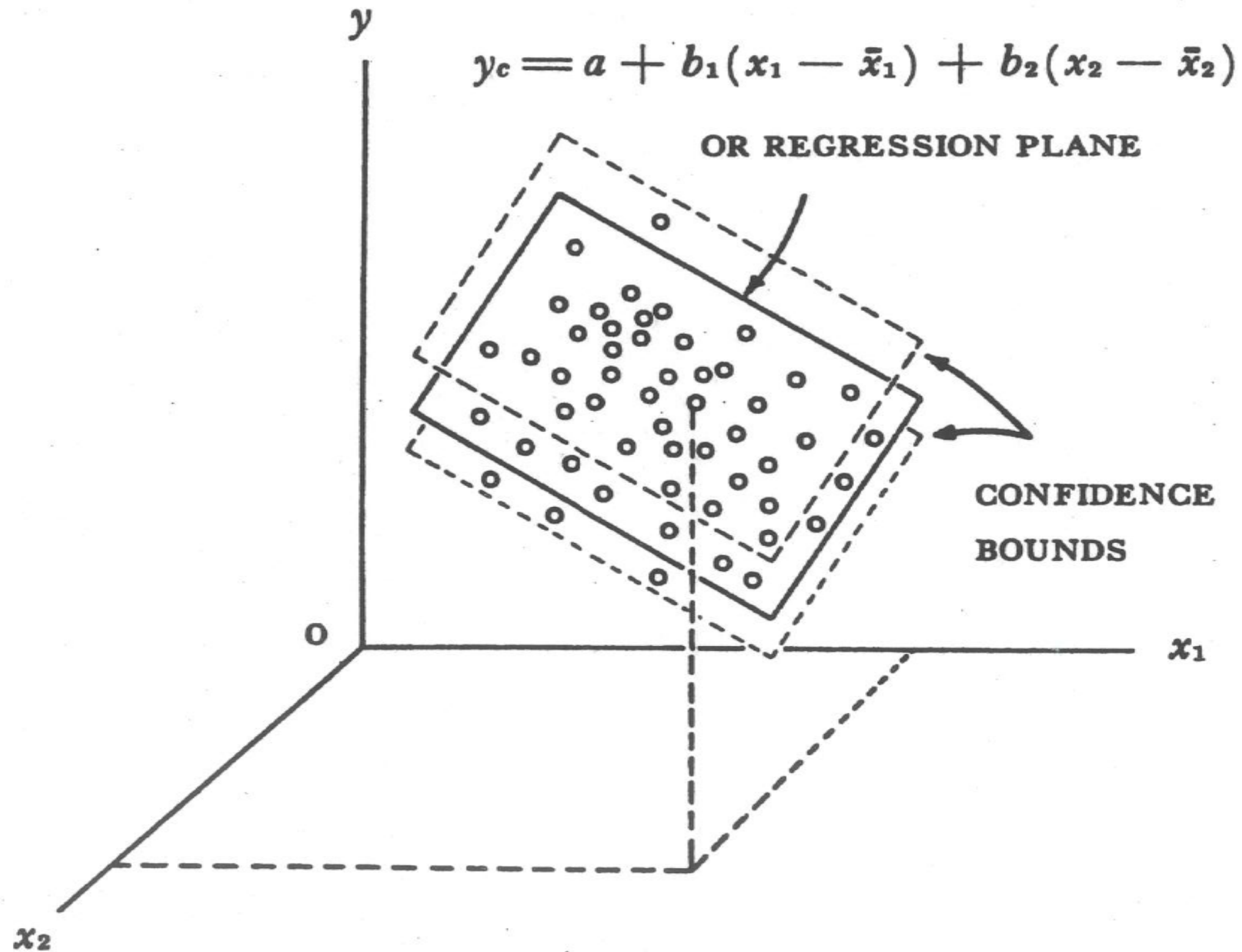


$$Y = X\beta + \mu \quad [Y_{m1}] = [X_{mm}] [\beta_{m1}] + [\mu_{m1}]$$

- Onde: Y_{m1} é um vetor de variáveis dependentes
- X_{mm} é uma matriz de regressores
- β_{m1} é um vetor de coeficientes e μ_{m1} é um vetor de resíduos
- Já que é igual a tudo que nós vimos, nada passa de um problema de otimização, basta encontrar os valores dos betas que minimizam o erro:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + e$$

$$e^2 = S_r = \sum_i^n (y_i \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni})^2$$



Regressão Linear Múltipla

- Avaliar as Variáveis Independentes individualmente X_1 , X_2 , etc.
- Criar gráficos de dispersão das Variáveis Independentes;
- Buscar redundancias;
- Tomar cuidado, pois o coeficiente de Determinação r^2 sempre aumentará quando utilizar mais variáveis. Utilizar o **Coeficiente de Determinação Ajustado**;
- Variáveis Colineares atrapalham o modelo;
- Variáveis **sem relação não ajudam** o modelo;
- Objetivo: Criar o melhor modelo com o menor número de variáveis.

Coeficiente de Determinação Ajustado

- Minimiza o aumento provocado pelo simples fato de adicionar mais variáveis ao modelo;

$$\bar{r}_a^2 = 1 - (1 - r^2) \left(\frac{n - 1}{n - (k + 1)} \right)$$

- n = quantidade de amostras
- k = quantidade de variáveis do modelo
- Assim pode escolher o modelo que tenha \bar{r}_a^2 Máximo com o menor número de variáveis;

- Da mesma forma que no modelo linear simples:
 - Requisitos
 - Linearidade entre as variáveis:
 - Normalmente Distribuídas
 - Pouca Colinearidade:
 - Residuais:
 - Distribuição Normal
 - Variância constante
 - Independentes

Modelos Lineares Generalizados

- “Uma parte importante de toda pesquisa de Modelagem estatística e envolve a procura de um modelo que seja o mais simples possível e que descreva bem os dados observados”. [Gauss e Clarisse, 2008]
- Porém em 1972 Nelder e Wedderburn propuseram o chamado Modelos Lineares Generalizados.
- Antes dos MLG utilizavam aproximações dos modelos lineares.
- Desde então diversos modelos foram propostos.
- Família exponencial:
$$f(x; \theta) = h(x) e^{\eta(\theta) t(x) - b(\theta)}$$

$$f(x; \theta) = h(x) e^{\eta(\theta) t(x) - b(\theta)}$$

$$f(x; \theta, \phi) = e^{\phi^{-1}[x\theta - b(\theta)]} + c(x, \phi)$$

- $b(\cdot)$ e $c(\cdot)$ são funções conhecidas
- ϕ é um componente aleatório > 0
- ϕ é um parâmetro de dispersão e seu inverso $\frac{1}{\phi}$ um parâmetro de precisão.

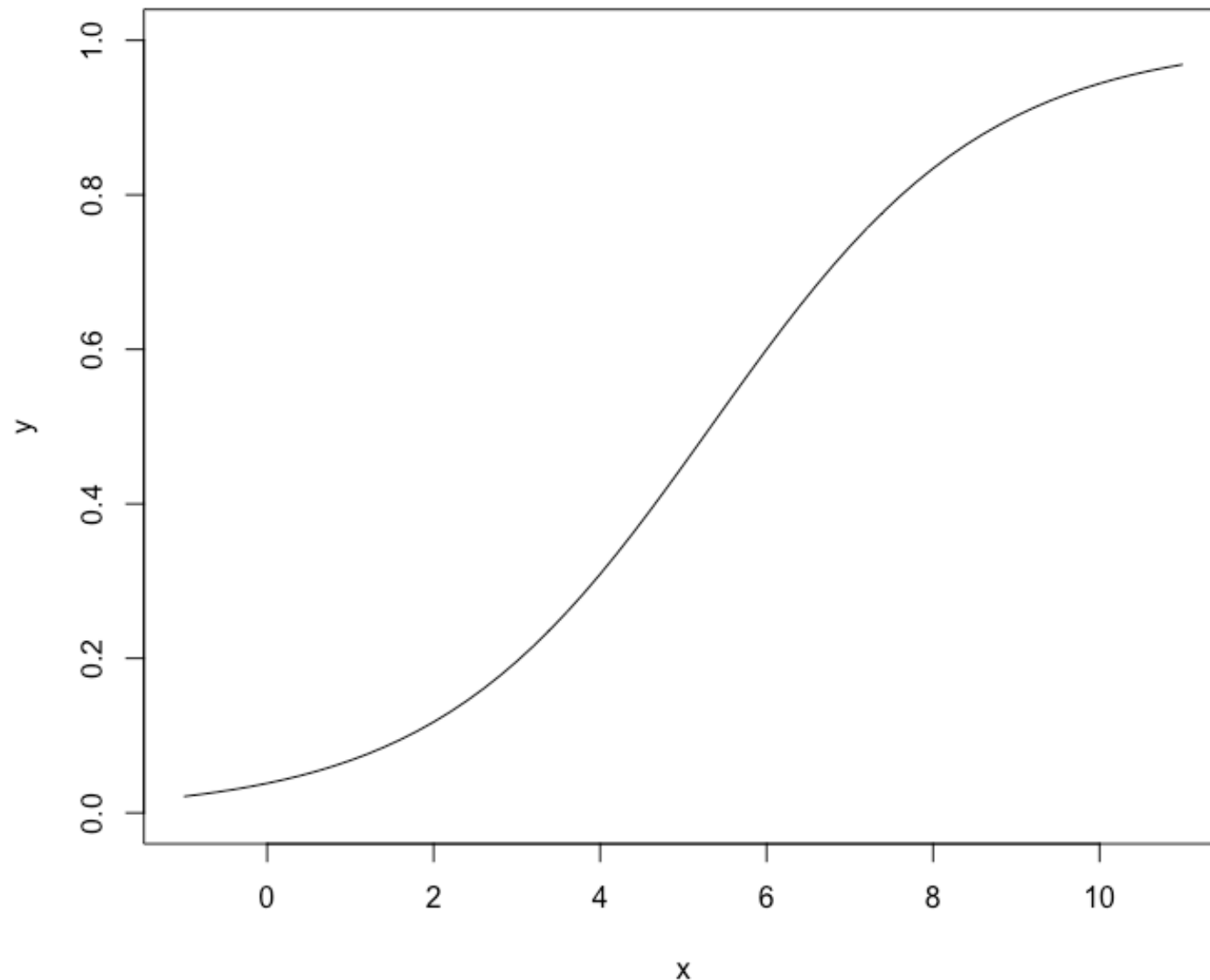
Modelos Lineares Generalizados

- **Normal**
 - **Distribuição de Gauss**
 - Modelo Linear Clássico
- **Binomial**
 - **Sucesso e Fracasso**
- **Poisson**
 - **Variáveis aleatórias discretas**
 - Contagens num intervalo de tempo
- **Gama**
 - **Dados assimétricos**
 - Precipitação de chuva, Cópia genética
- **Gaussiano Inverso**
 - **Dados muito assimétricos**
 - Tempo de vida

Família exponencial

Distribuição	ϕ	θ	$b(\theta)$	$c(y, \phi)$	$\mu(\theta)$	$V(\mu)$
Normal: $N(\mu, \sigma^2)$	σ^2	μ	$\frac{\theta^2}{2}$	$-\frac{1}{2} \left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right]$	θ	1
Poisson: $P(\mu)$	1	$\log \mu$	e^θ	$-\log y!$	e^θ	μ
Binomial: $B(m, \pi)$	1	$\log \left(\frac{\mu}{m - \mu} \right)$	$m \log(1 + e^\theta)$	$\log \binom{m}{y}$	$\frac{me^\theta}{1 + e^\theta}$	$\frac{\mu}{m}(m - \mu)$
Binomial Negativa: $BN(\mu, k)$	1	$\log \left(\frac{\mu}{\mu + k} \right)$	$-k \log(1 - e^\theta)$	$\log \left[\frac{\Gamma(k + y)}{\Gamma(k)y!} \right]$	$k \frac{e^\theta}{1 - e^\theta}$	$\mu \left(\frac{\mu}{k} + 1 \right)$
Gama: $G(\mu, \nu)$	ν^{-1}	$-\frac{1}{\mu}$	$-\log(-\theta)$	$\nu \log(\nu y) - \log y - \log \Gamma(\nu)$	$-\frac{1}{\theta}$	μ^2
Normal Inversa: $IG(\mu, \sigma^2)$	σ^2	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{1/2}$	$-\frac{1}{2} \left[\log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y} \right]$	$(-2\theta)^{-1/2}$	μ^3

Regressão Binomial



$$f(x; \theta) = \binom{m}{x} \theta^x (1 - \theta)^{m-x} = \frac{m!}{x! (m-x)!} \theta^x (1 - \theta)^{m-x}$$

Regressão Binomial

- É uma regressão que mede a **probabilidade** de que um certo evento ocorra dados n elementos.
- Portanto deve prever um valor entre 0 e 1 ($0 < p < 1$).
- Sucesso ou Fracasso
- Os ensaios são independentes
- Média $n \cdot p$ e variância $np(1-p)$
- Média é o valor esperado em n vezes
- Suporte $0(1)n / n$

Binomial

- Ensaios do tipo Dose-Resposta
 - Dose de medicamentos
 - Inseticidas
- Regressões Logísticas
 - Alunos aprovados ou reprovados
 - Doença ou Estado Clínico

Binomial: Logística

$$p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

$$\frac{p}{1 - p} = \frac{\text{probabilidade sucesso}}{\text{probabilidade fracasso}} = e^{\alpha + \beta x}$$

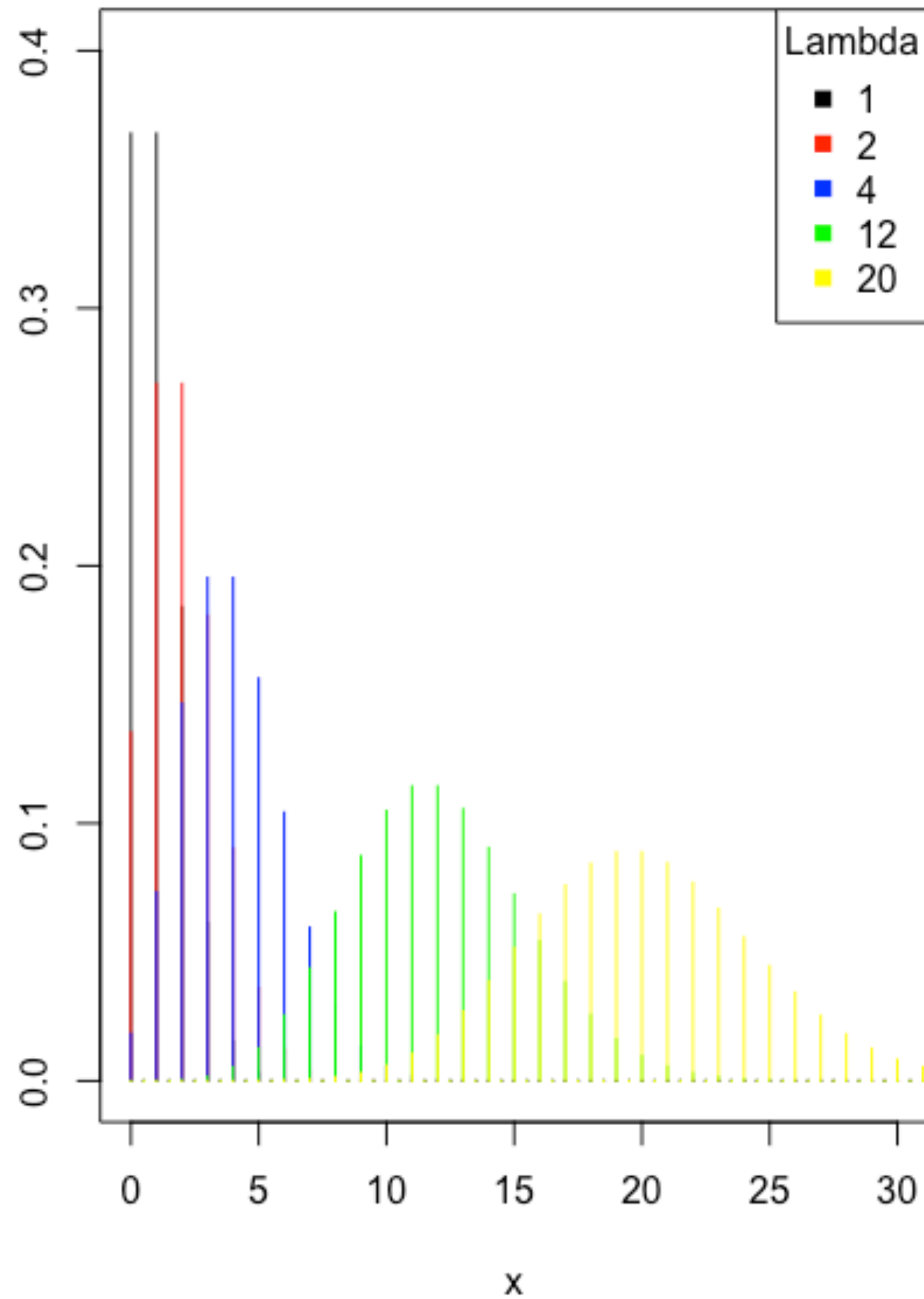
$$\ln \left[\frac{p}{1 - p} \right] = \alpha + \beta x$$

$$P(\lambda) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Poisson

- Utilizada para valores **discretos**
- Possui apenas um parâmetro (λ)
- Variância é igual à média
- A taxa média de ocorrência é constante no tempo
- Os intervalos disjuntos são independentes. Ou seja, cada intervalo não possui relação com os outros;
- Utilizado para contar número de eventos no período do tempo
- Suporte 0(1) até Infinito

Poisson



Poisson

- Defeitos por unidade de tempo
- Defeitos por unidade de área
- Acessos por unidade de tempo
- Em outras palavras, dados de contagens por unidades

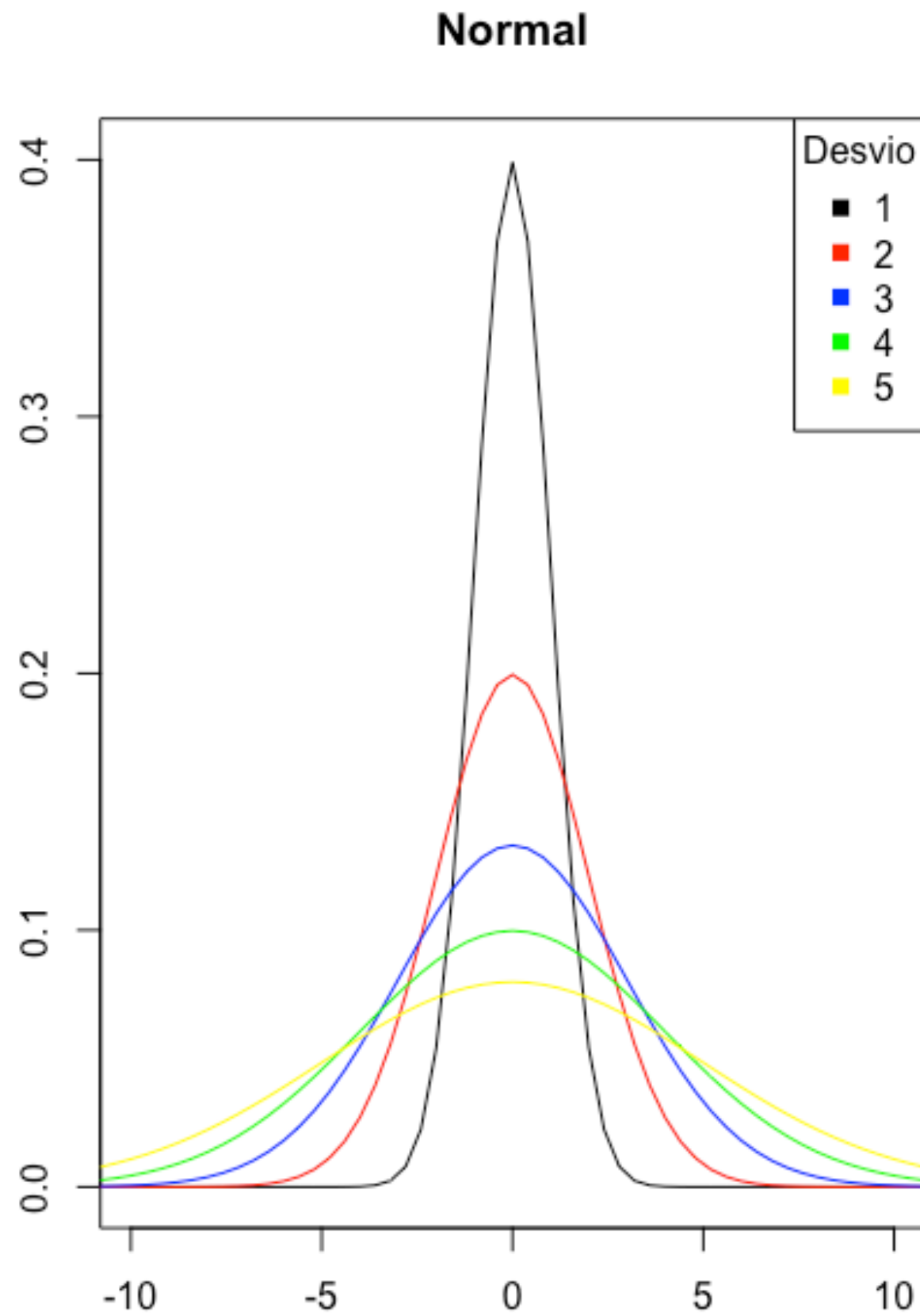
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Gaussiana

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$X \sim N(\mu, \sigma)$$

Gaussiana

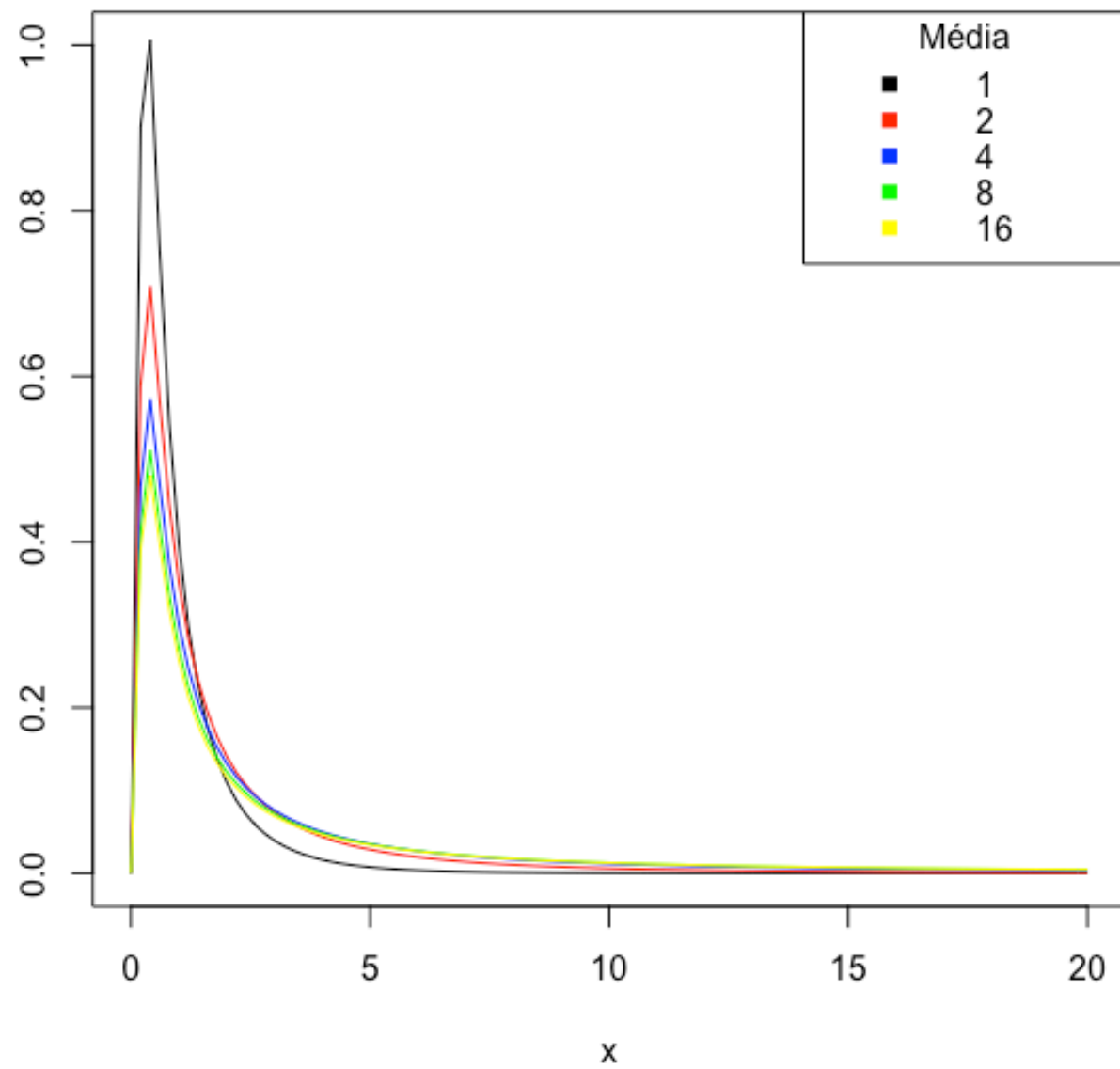


- Modelos contínuos
- Suporte - infinito até + infinito

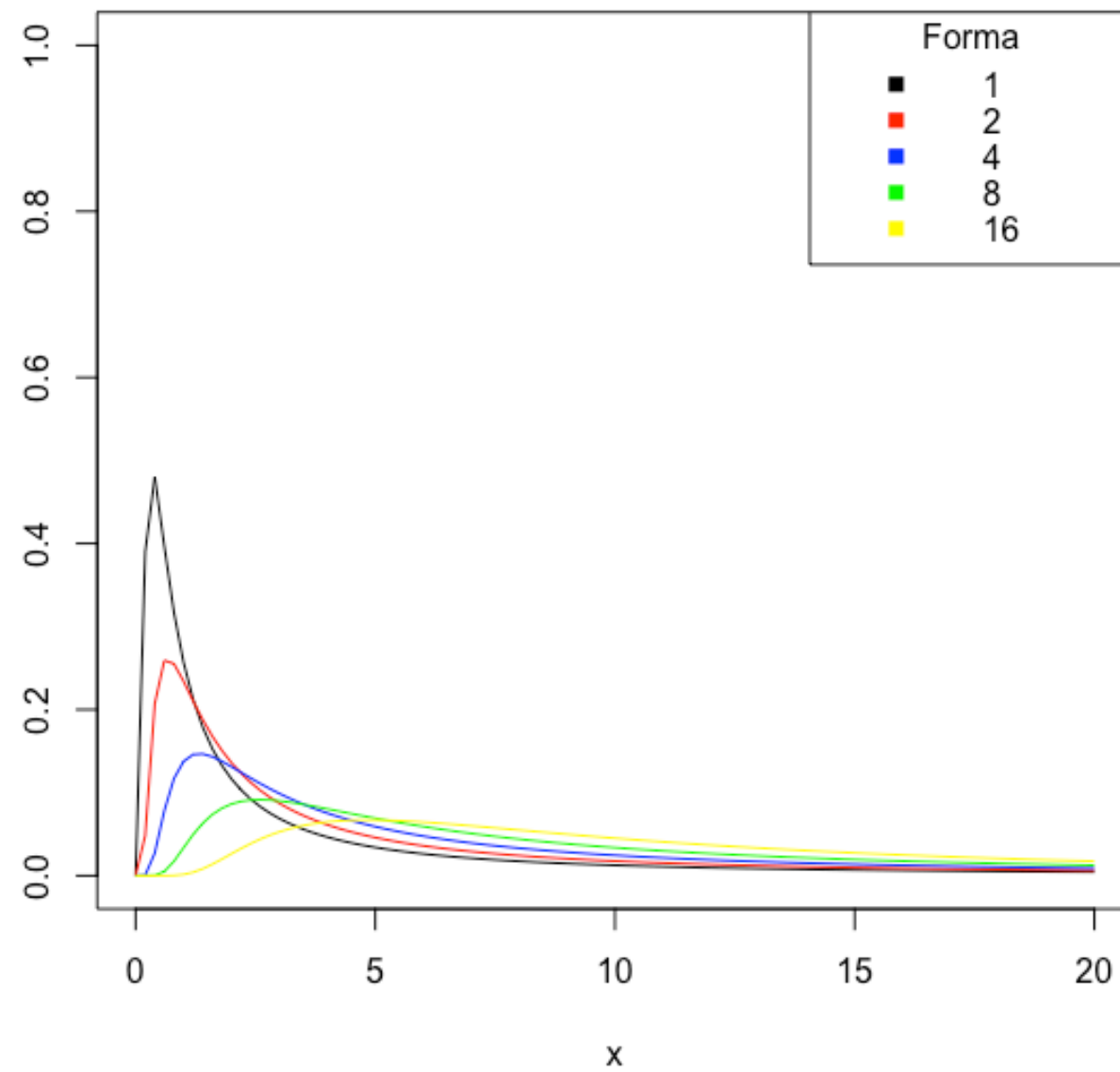
Gaussiana Inversa

- Útil para estudos de análise de regressão com dados assimétricos.
- Útil para tempo de vida
- Início em 0
- Biparamétrica

Gaussiana Inversa



Gaussiana Inversa



$$f(y_i; \mu, \phi) = \frac{1}{\Gamma(\phi)} \left(\frac{\phi y}{\mu} \right)^{\phi} \exp \left(-\frac{\phi y}{\mu} \right) d(\log y)$$

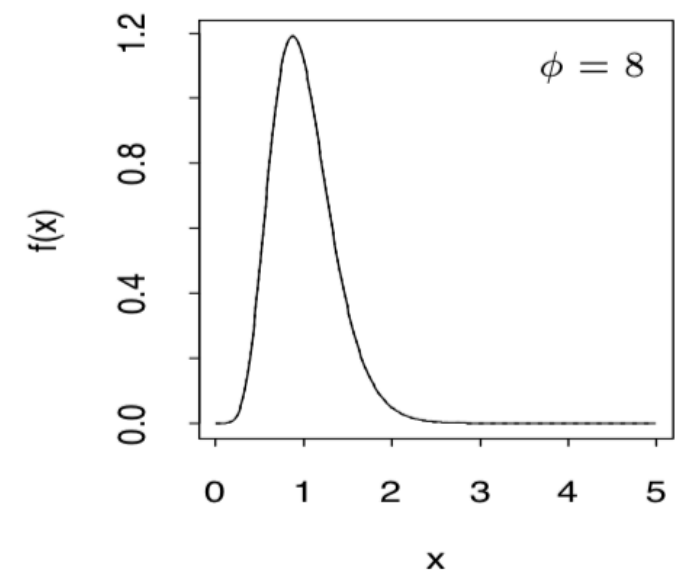
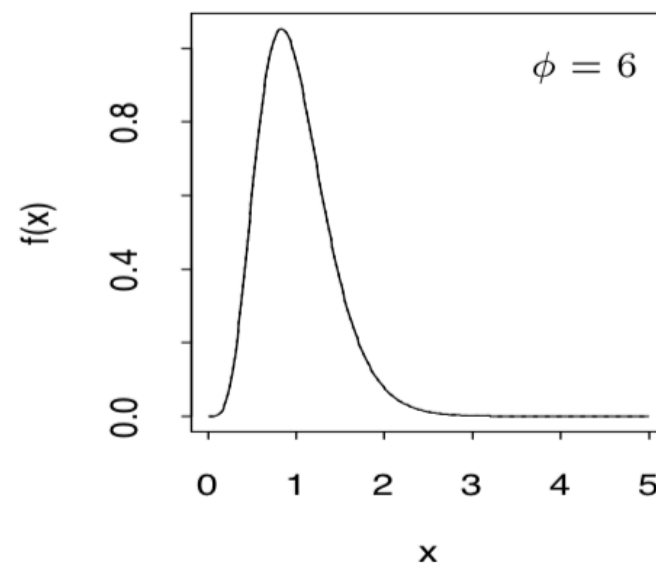
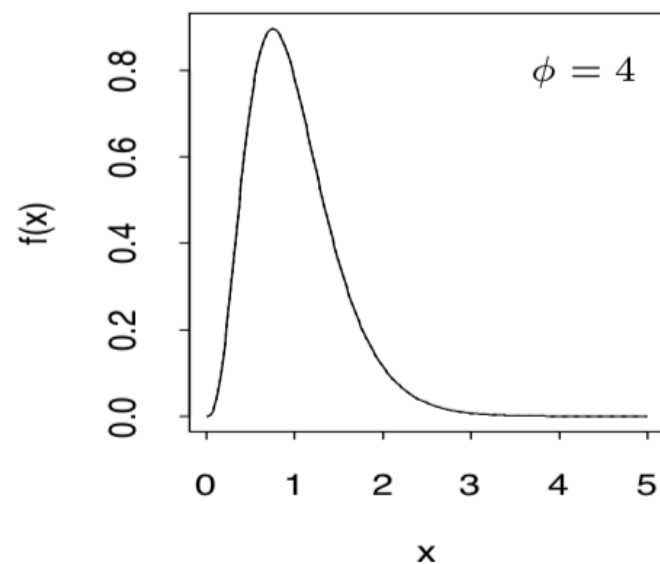
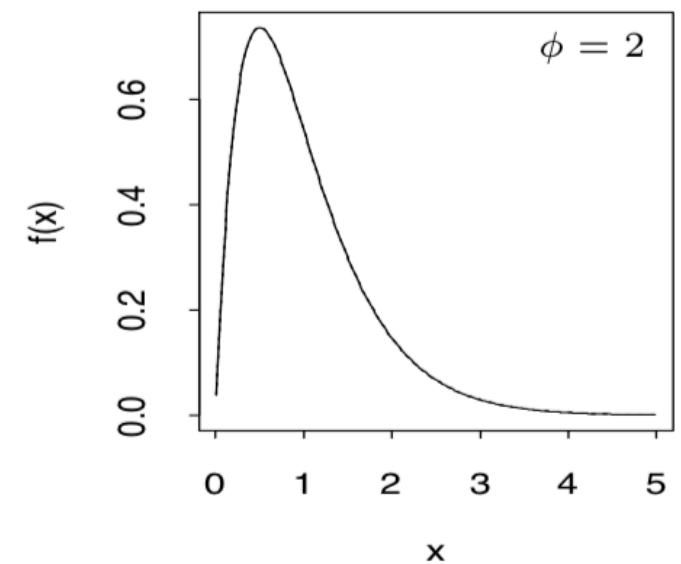
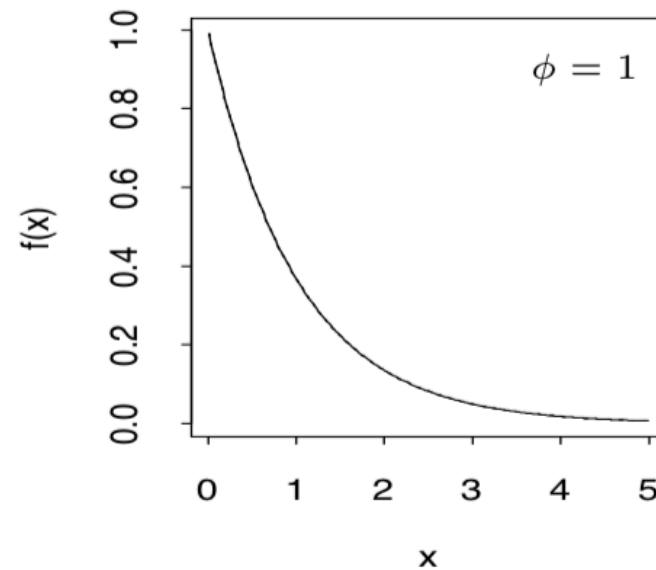
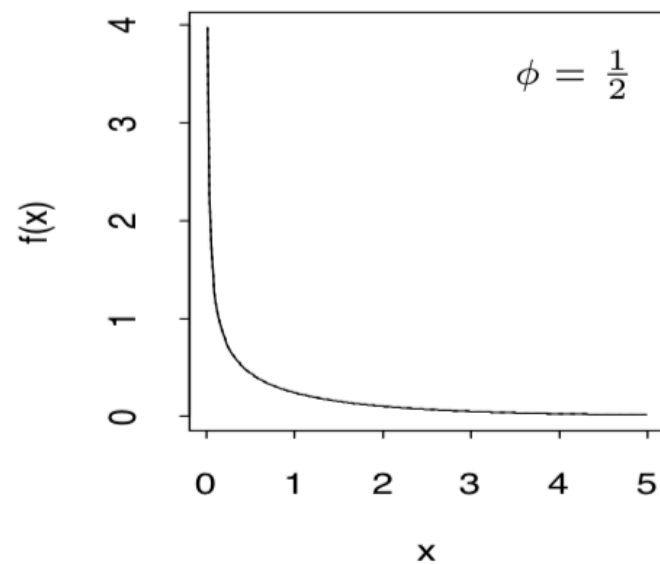
$$y > 0, \phi > 0, \mu > 0, \Gamma(\phi) = \int_0^{\infty} t^{\phi-1} e^{-t} dt$$

Gama

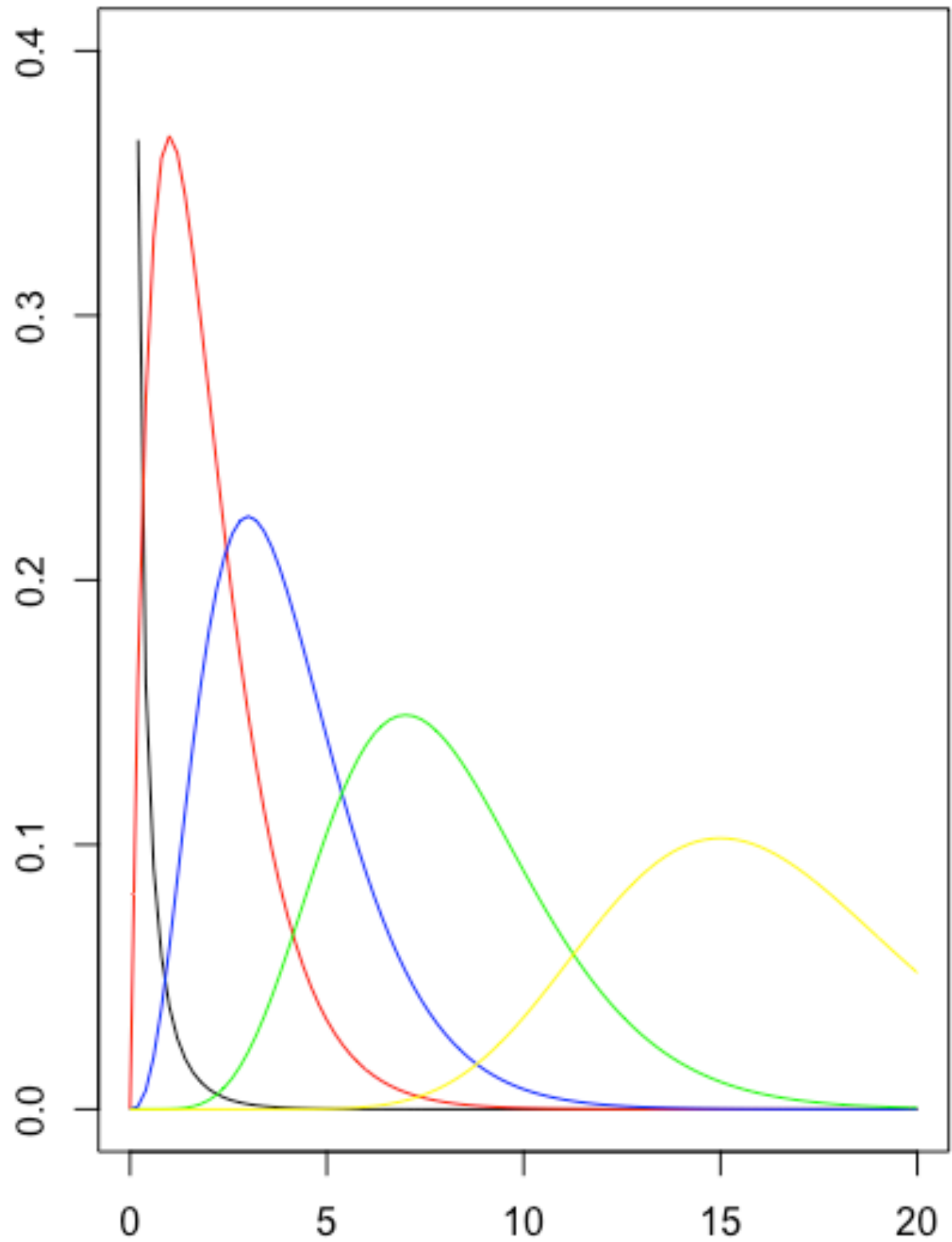
- Variáveis de resposta são independentes
- Modelo de resposta contínua
- Variável resposta assume valores positivos
- Dados positivos bastante assimétricos
- Quanto mais ϕ aumenta mais a distribuição fica simétrica à média
- Suporte de 0 até infinito

Gama

$$f(y_i; \mu, \phi) = \frac{1}{\Gamma(\phi)} \left(\frac{\phi y}{\mu} \right)^\phi \exp \left(-\frac{\phi y}{\mu} \right) d(\log y) \quad y > 0, \phi > 0, \mu > 0, \Gamma(\phi) = \int_0^\infty t^{\phi-1} e^{-t} dt$$



gama



Gama

- Dados assimétricos
- Análise de sobrevivência
- Análises onde parte dos dados não foram observadas falhas (censura)

- "Uma das vantagens de usarmos ligações canônicas é que as mesmas garantem a concavidade de $L(b)$ e consequentemente muitos resultados assintóticos são obtidos mais facilmente." (Gilberto A. Paula, 2010)

Distribuição	Normal	Binomial	Poisson	Gama	N. Inversa
Ligação	$\mu = \eta$	$\log \left\{ \frac{\mu}{1-\mu} \right\} = \eta$	$\log \mu = \eta$	$\mu^{-1} = \eta$	$\mu^{-2} = \eta$

No R

Tabela 1.5

Quantidades úteis para diagnóstico obtidas no R.

Símbolo	Descrição	Função	Elemento
\mathbf{h}	Alavanca	lm.influence()	hat
$\hat{\boldsymbol{\beta}}$	Coeficientes	coef()	
\mathbf{r}	Resíduos	resid()	
s	Desvio padrão amostral	summary()	sigma
$\mathbf{s}_{(i)}$	Desvio padrão sem observação i	lm.influence()	sigma
$\hat{\boldsymbol{\beta}}_{(i)}$	Coeficiente sem observação i	lm.influence()	coef
$(\mathbf{X}^T \mathbf{X})^{-1}$	Covariância de $\hat{\boldsymbol{\beta}}$ sem s^2	summary()	cov.unscaled