

CE-092

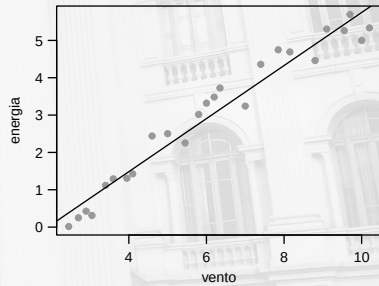
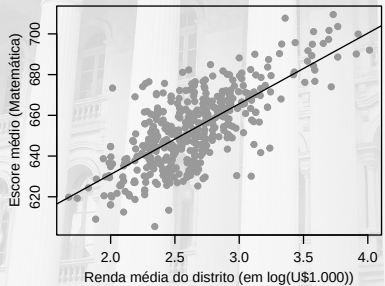
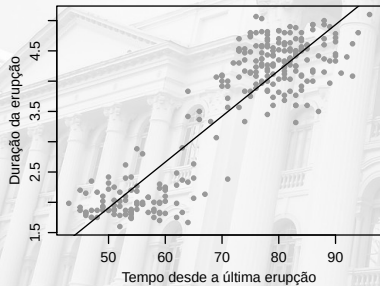
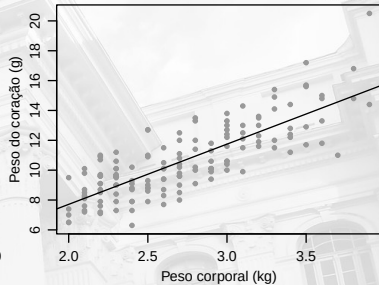
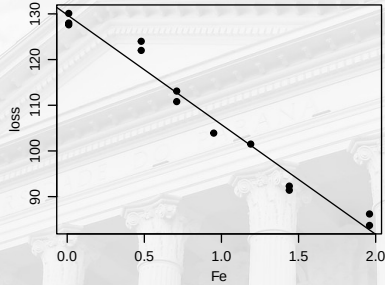
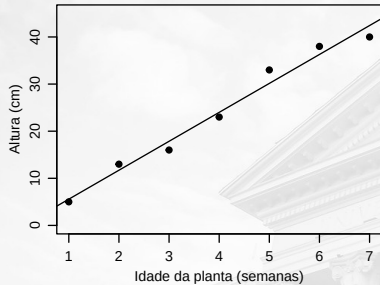
Modelos de regressão linear e extensões

Paulo Justiniano Ribeiro Jr

Curso de Estatística e Ciência de Dados
Departamento de Estatística
Setor de Ciências Exatas
Universidade Federal do Paraná



Exemplos - motivação para regressão



Regressão linear simples

- Semelhanças e diferenças nos exemplos

Regressão linear simples

- ▶ Semelhanças e diferenças nos exemplos
- ▶ Retas de "regressão" nos exemplos:

Regressão linear simples

- ▶ Semelhanças e diferenças nos exemplos
- ▶ Retas de "regressão" nos exemplos:
 - ▶ Como são obtidas?

Regressão linear simples

- ▶ Semelhanças e diferenças nos exemplos
- ▶ Retas de "regressão" nos exemplos:
 - ▶ Como são obtidas?
 - ▶ Como são interpretadas?

Regressão linear simples

- ▶ Semelhanças e diferenças nos exemplos
- ▶ Retas de "regressão" nos exemplos:
 - ▶ Como são obtidas?
 - ▶ Como são interpretadas?
 - ▶ Como avaliar se descrevem bem os dados?

Regressão linear simples

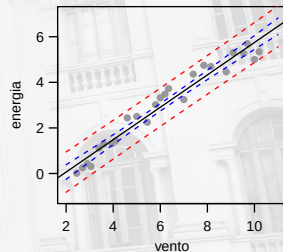
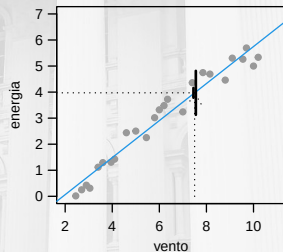
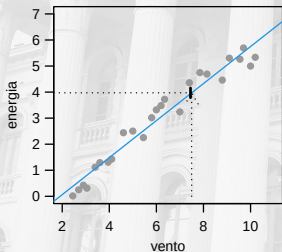
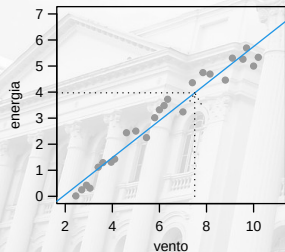
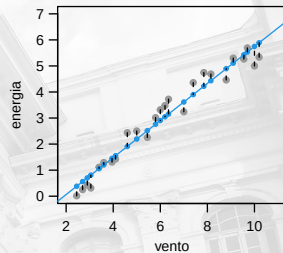
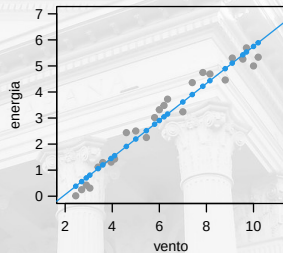
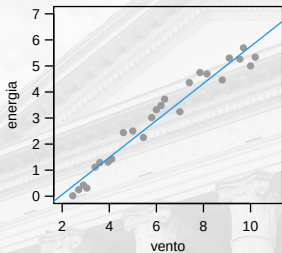
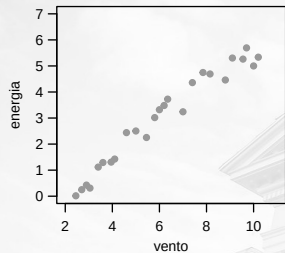
- ▶ Semelhanças e diferenças nos exemplos
- ▶ Retas de "regressão" nos exemplos:
 - ▶ Como são obtidas?
 - ▶ Como são interpretadas?
 - ▶ Como avaliar se descrevem bem os dados?
 - ▶ Como avaliar a incerteza?

Regressão linear simples

- ▶ Semelhanças e diferenças nos exemplos
- ▶ Retas de "regressão" nos exemplos:
 - ▶ Como são obtidas?
 - ▶ Como são interpretadas?
 - ▶ Como avaliar se descrevem bem os dados?
 - ▶ Como avaliar a incerteza?
 - ▶ Como avaliar se "valem a pena"?

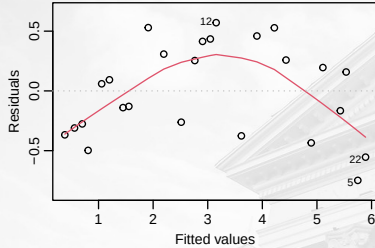
- ▶ Semelhanças e diferenças nos exemplos
- ▶ Retas de "regressão" nos exemplos:
 - ▶ Como são obtidas?
 - ▶ Como são interpretadas?
 - ▶ Como avaliar se descrevem bem os dados?
 - ▶ Como avaliar a incerteza?
 - ▶ Como avaliar se "valem a pena"?
 - ▶ Como utilizar para fazer inferência e predição?

Exemplo energia

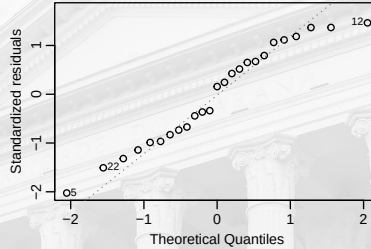


Diagnósticos

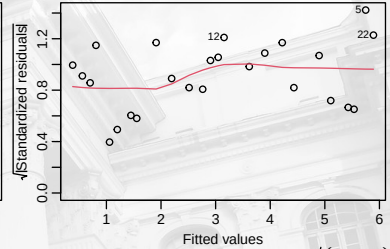
Residuals vs Fitted



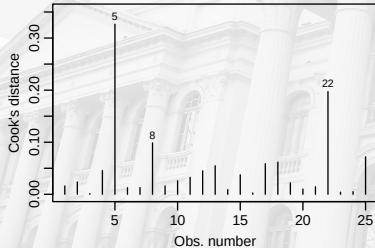
Q-Q Residuals



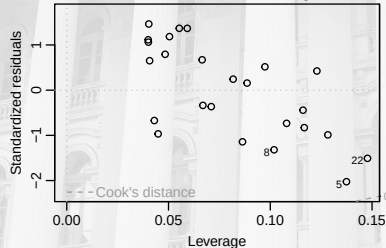
Scale-Location



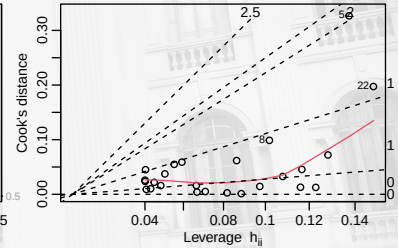
Cook's distance



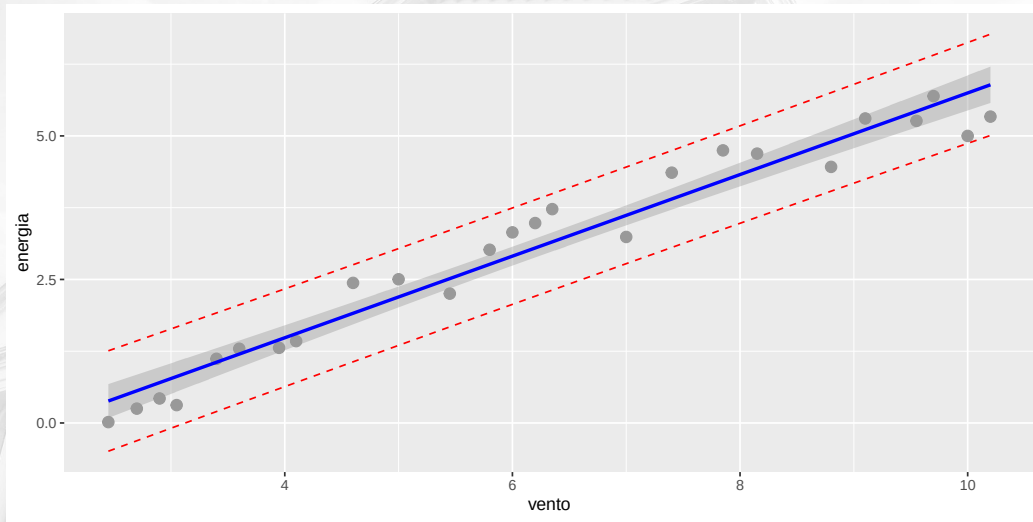
Residuals vs Leverage



Cook's dist vs Leverage $h_{ii}/(1-h_{ii})$



Bandas de confiança e predição



Resultados

```
ajuste <- lm(energia ~ vento, data = eolica)
coef(ajuste)

## (Intercept)      vento
## -1.3585119    0.7106607

confint(ajuste)

##              2.5 %      97.5 %
## (Intercept) -1.7978782 -0.9191455
## vento       0.6442299  0.7770915

anova(ajuste)

## Analysis of Variance Table
##
## Response: energia
##      Df Sum Sq Mean Sq F value    Pr(>F)
## vento  1  77.551   77.551   489.74 < 2.2e-16 ***
## Residuals 23   3.642    0.158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ajuste)

##
## Call:
## lm(formula = energia ~ vento, data = eolica)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74840 -0.31027  0.05951  0.30793  0.57072
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.35851     0.21239  -6.396 1.58e-06 ***
## vento        0.71066     0.03211  22.130 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3979 on 23 degrees of freedom
## Multiple R-squared:  0.9551, Adjusted R-squared:  0.9532
## F-statistic: 489.7 on 1 and 23 DF,  p-value: < 2.2e-16
```

- O modelo de **regressão linear simples** é usualmente escrito da forma:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

- Como a regressão diz respeito à distribuição $[Y|X = x]$, o modelo define

$$E[Y_i|X = x_i] = \beta_0 + \beta_1 x_i$$

- Vamos relacionar com o que já vimos até aqui e a especificação mais geral.

Modelos de regressão (linear, simples e gaussiana) para os exemplos

$$Y_{ij}|x_i \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \mu_i = \beta_0 + \beta_1 x_i$$

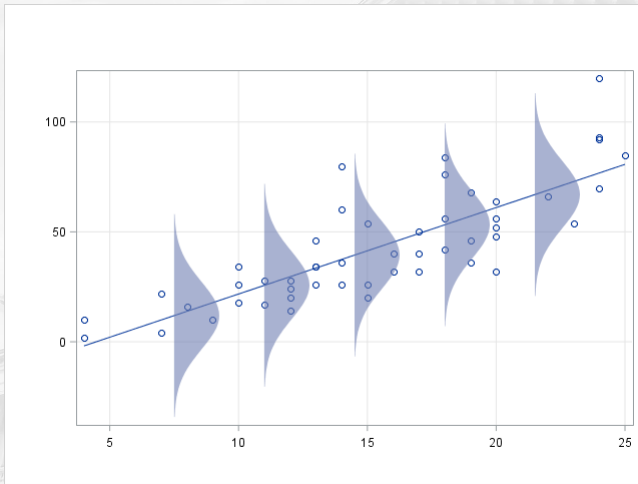
em que

Y_{ij} : observação j correspondente ao valor de x_i

$\mu_{ij} = \mu_i = \beta_0 + \beta_1 x_i \longrightarrow$ para mesmo x_i tem-se a mesma média

$\sigma^2 \longrightarrow$ todas observações possuem a mesma variância

Visualizando o modelo de regressão linear simples



$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \mu_i = \beta_0 + \beta_1 x_i$$

Figura 1. Regressão linear simples

Modelagem estatística e regressão

Estrutura em comum:

- ▶ Variável Y : resposta ou **dependente** (*response, outcome, output or dependent*)
- ▶ Uma ou mais variável $X_1, X_2, \dots, X_p$: preditor, explicativa, **independente**, covariável (*predictor, input, independent, explanatory or covariate*)

Modelagem estatística e regressão

Estrutura em comum:

- ▶ Variável Y : resposta ou **dependente** (*response, outcome, output or dependent*)
- ▶ Uma ou mais variável $X_1, X_2, \dots, X_p$: preditor, explicativa, **independente**, covariável (*predictor, input, independent, explanatory or covariate*)

Forma básica:

Objetivo: explicar Y a partir de alguma função dos X .

$$Y = f(X) + \epsilon$$

Modelagem estatística e regressão

Estrutura em comum:

- ▶ Variável Y : resposta ou **dependente** (*response, outcome, output or dependent*)
- ▶ Uma ou mais variável $X_1, X_2, \dots, X_p$: preditor, explicativa, **independente**, covariável (*predictor, input, independent, explanatory or covariate*)

Forma básica:

Objetivo: explicar Y a partir de alguma função dos X .

$$Y = f(X) + \epsilon$$

Forma mais geral:

Objetivo: explicar os parâmetros da distribuição de Y a partir de alguma função dos X .

$$Y_{ij}^{(\lambda)} \sim \text{Dist}(\mu_{ij} = E[Y_{ij}], \sigma_{ij}^2 = \text{Var}[Y_{ij}])$$

$$g_1(\mu_{ij}) = f_1(x_{ij}, \beta)$$

$$g_2(\sigma_{ij}^2) = f_2(z_{ij}, \varphi)$$

Mas afinal, de onde vem os modelos de regressão?

- A teoria (física) pode sugerir o modelo. Por exemplo, segundo a lei de Ohm a voltagem aplicada nos terminais de um condutor é proporcional à corrente elétrica que o percorre. Logo, a relação entre as variáveis é linear, induzindo o modelo;

Mas afinal, de onde vem os modelos de regressão?

- ▶ A teoria (física) pode sugerir o modelo. Por exemplo, segundo a lei de Ohm a voltagem aplicada nos terminais de um condutor é proporcional à corrente elétrica que o percorre. Logo, a relação entre as variáveis é linear, induzindo o modelo;
- ▶ Experiência empírica. Se um particular modelo proporcionou bom ajuste aos dados em estudos similares, possivelmente ele vai se ajustar bem aos dados do presente estudo;

Mas afinal, de onde vem os modelos de regressão?

- ▶ A teoria (física) pode sugerir o modelo. Por exemplo, segundo a lei de Ohm a voltagem aplicada nos terminais de um condutor é proporcional à corrente elétrica que o percorre. Logo, a relação entre as variáveis é linear, induzindo o modelo;
- ▶ Experiência empírica. Se um particular modelo proporcionou bom ajuste aos dados em estudos similares, possivelmente ele vai se ajustar bem aos dados do presente estudo;
- ▶ Não há qualquer modelo indicado a priori. A escolha de um modelo pode resultar da exploração dos dados e comparação de diferentes especificações.
- ▶ A relação linear do preditor pode ser substituída por equações não lineares (reg. não linear), funções suavizadas (reg. splines), dentre outras generalizações.

Taylor!

Série de Taylor:

Uma função pode ser aproximada por um polinômio ao redor de um ponto $x = a$,

$$f(x) \approx f(a) + f'(a)(x - a) + f''(a)\frac{(x - a)^2}{2!} + f'''(a)\frac{(x - a)^3}{3!} + \dots + f^{(n)}(a)\frac{(x - a)^n}{n!}.$$

Taylor!

Série de Taylor:

Uma função pode ser aproximada por um polinômio ao redor de um ponto $x = a$,

$$f(x) \approx f(a) + f'(a)(x - a) + f''(a)\frac{(x - a)^2}{2!} + f'''(a)\frac{(x - a)^3}{3!} + \dots + f^{(n)}(a)\frac{(x - a)^n}{n!}.$$

Um modelo de regressão mais geral seria da forma

$$y = f(x) + \epsilon.$$

Modelo de regressão como aproximação de 1ª ordem ao redor de $\bar{x}$:

$$f(x) \approx f(\bar{x}) + f'(\bar{x})(x - \bar{x}).$$

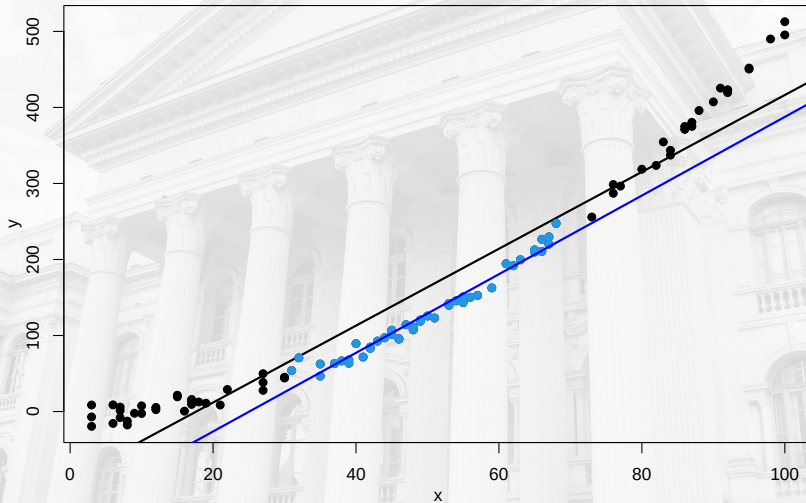
Que pode ser reescrito como:

$$f(x) \approx \beta_0 + \beta_1 x,$$

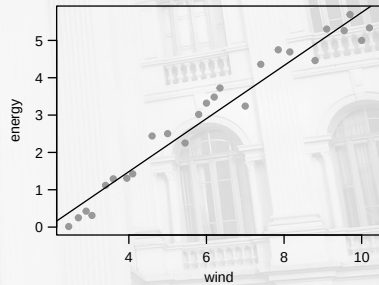
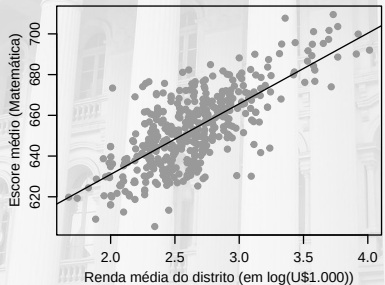
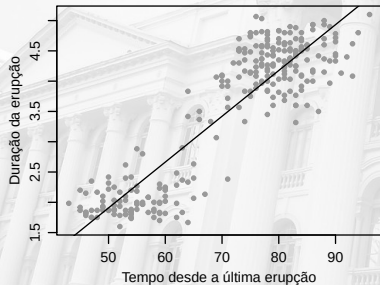
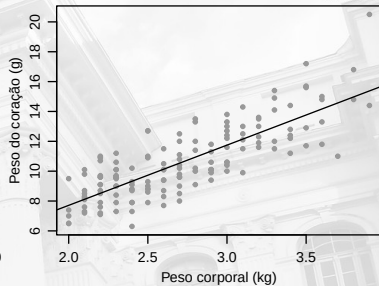
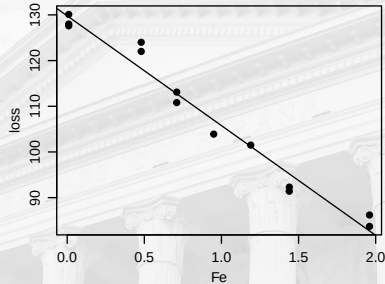
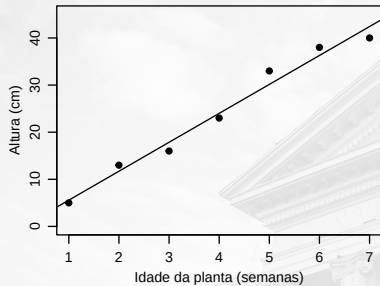
em que

$$\beta_0 = f(\bar{x}) - f'(\bar{x})\bar{x} \quad \text{e} \quad \beta_1 = f'(\bar{x}).$$

Regressão como aproximação local



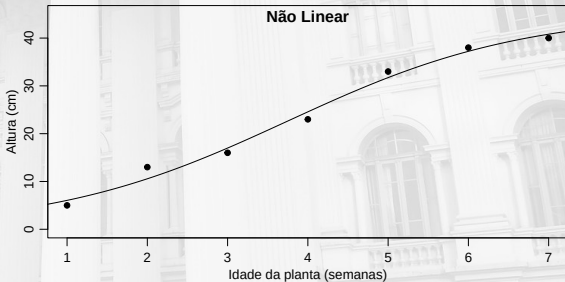
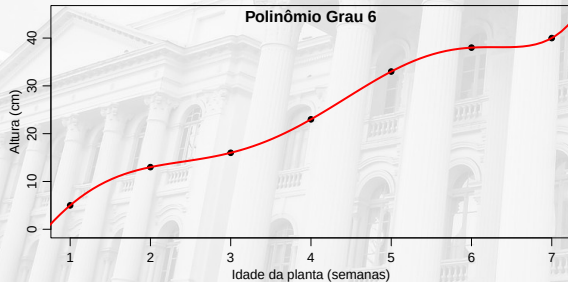
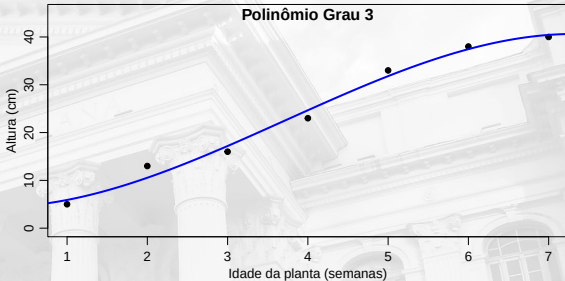
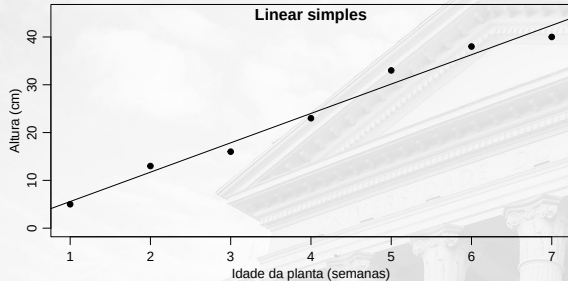
Exemplos - motivação para regressão



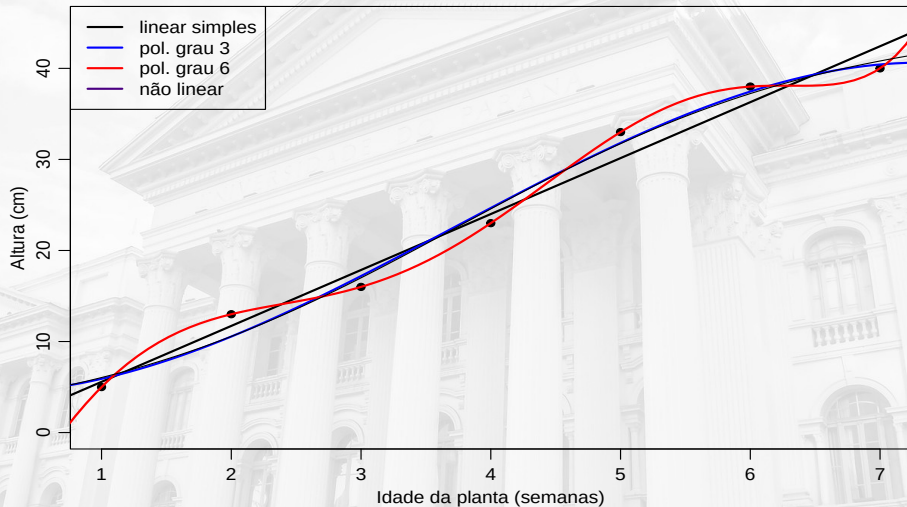
Regressão linear simples

- ▶ Semelhanças e diferenças nos exemplos
- ▶ Retas de "regressão" nos exemplos:
 - ▶ Como são obtidas?
 - ▶ Como são interpretadas?
 - ▶ Como avaliar se descrevem bem os dados?
 - ▶ Como avaliar a incerteza?
 - ▶ Como avaliar se "valem a pena"?
 - ▶ Como utilizar para fazer inferência e predição?
- ▶ Os modelos e ajustes podem ser "melhorados"?

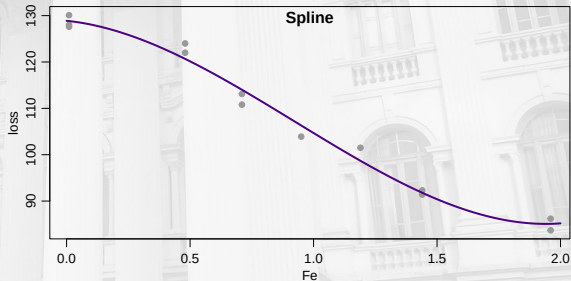
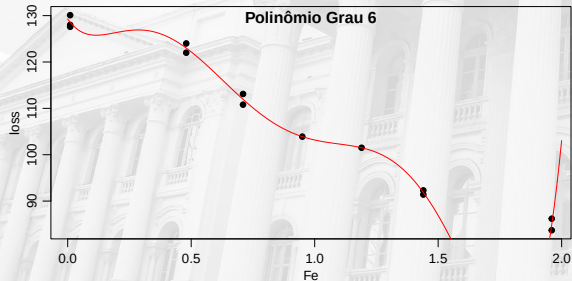
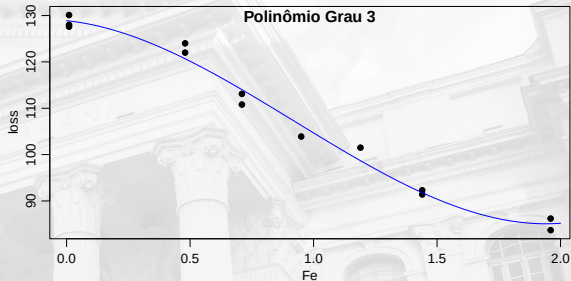
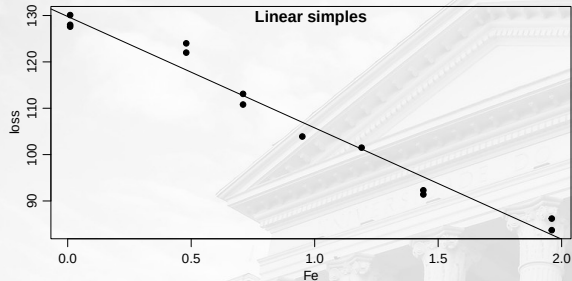
Motivação para regressão: exemplo crescimento de planta



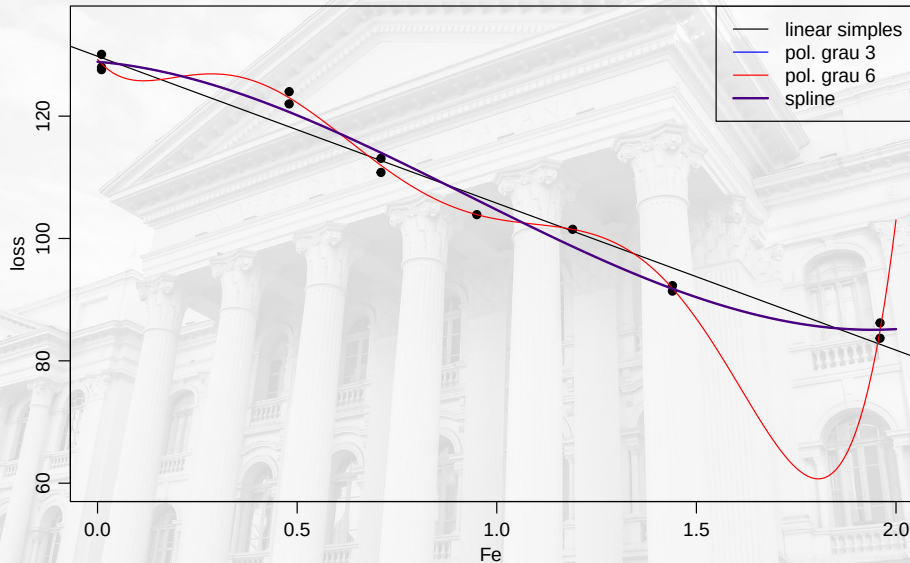
Motivação para regressão: exemplo crescimento de planta



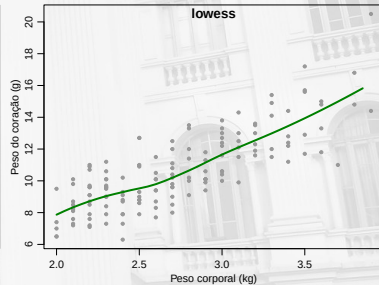
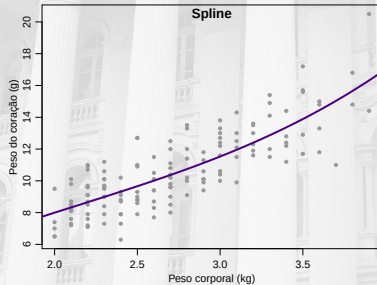
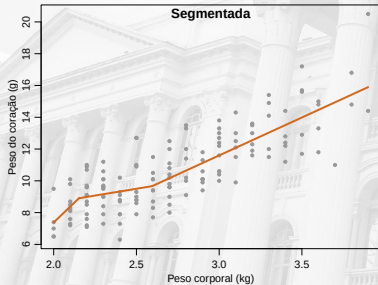
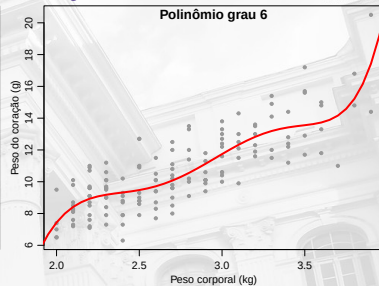
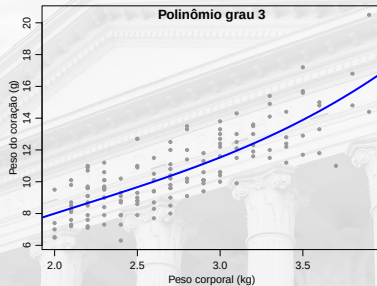
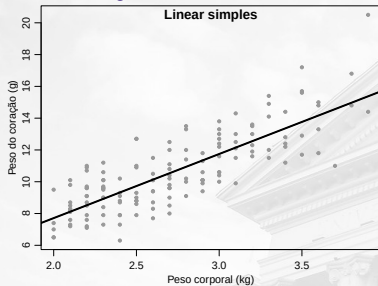
Motivação para regressão: exemplo perda de material



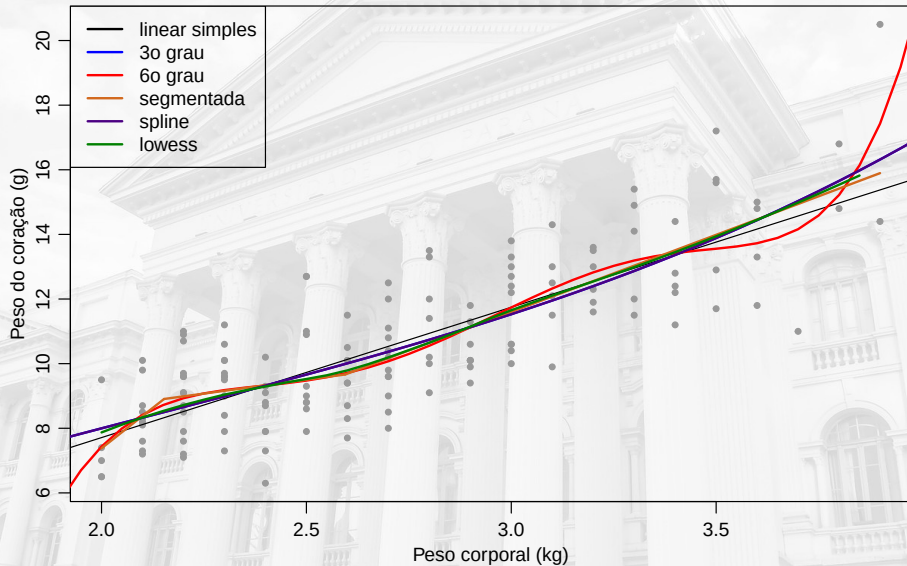
Motivação para regressão: exemplo perda de material



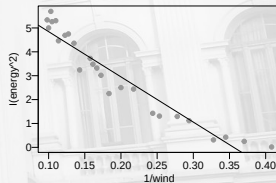
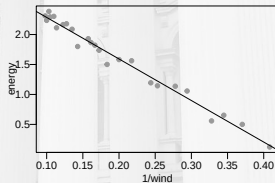
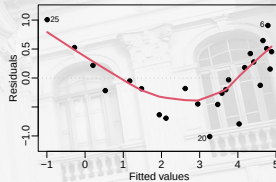
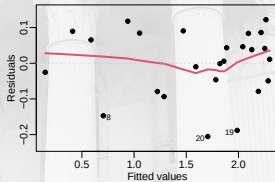
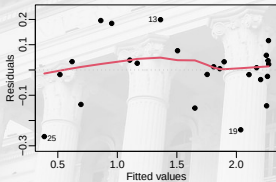
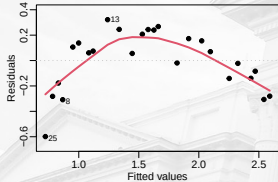
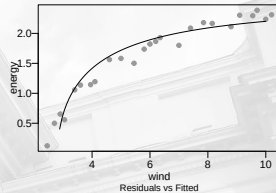
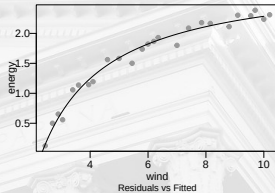
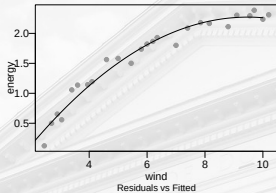
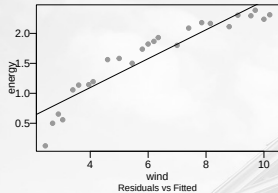
Motivação para regressão: exemplo peso do coração



Motivação para regressão: exemplo peso do coração



Motivação para regressão: exemplo energia



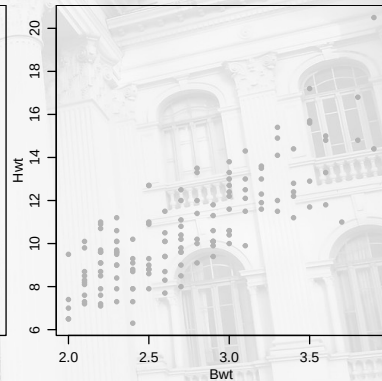
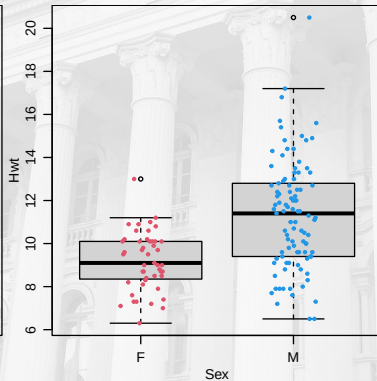
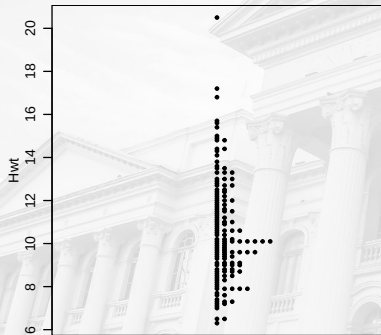
Tudo junto e misturado: dados **cats**

Descritiva, Probabilidades, Inferência, Modelagem

Variável de interesse: peso do coração de animais (gatos)

Informações acessórias: sexo e peso corporal dos animais

O que se pode *inferir* sobre o peso do coração dos animais?



Exemplo cats: estimando a média

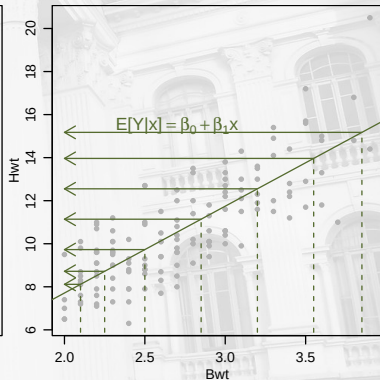
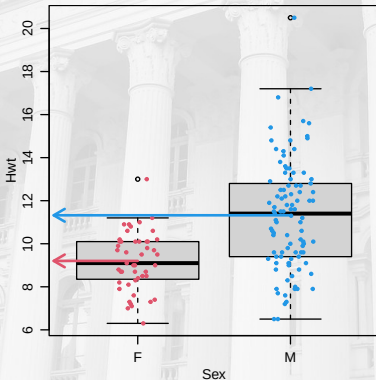
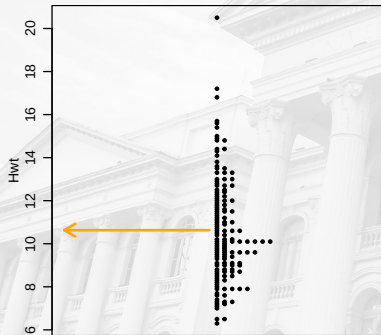
Descritiva, Probabilidades, Inferência, Modelagem

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \mu$$

$$\mu_{ij} = \mu_i, i \in \{F, M\}$$

$$\mu_{ij} = \mu_j = \beta_0 + \beta_1 x_j$$



Estimação

Modelos podem ser escritos como:

Modelo 0 : $Y_j = \mu + \epsilon_j$

parâmetros: μ e σ

Modelo 1 : $Y_{ij} = \mu_i + \epsilon_{ij}$

parâmetros: μ_1, μ_2 e σ

Modelo 2 : $Y_j = \beta_0 + \beta_1 x_j + \epsilon_j$

parâmetros: β_0, β_1 e σ

Estimação

Modelos podem ser escritos como:

Modelo 0 : $Y_j = \mu + \epsilon_j$

parâmetros: μ e σ

Modelo 1 : $Y_{ij} = \mu_i + \epsilon_{ij}$

parâmetros: μ_1, μ_2 e σ

Modelo 2 : $Y_j = \beta_0 + \beta_1 x_j + \epsilon_j$

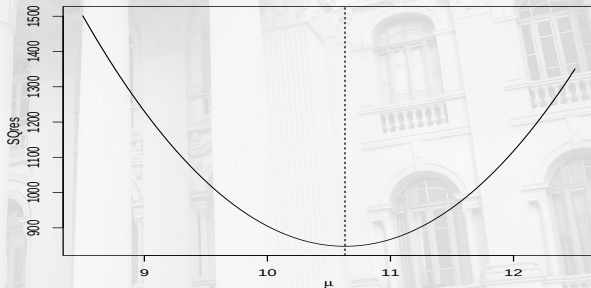
parâmetros: β_0, β_1 e σ

Parâmetros podem ser estimados por mínimos quadrados, por exemplo no Modelo 1:

$$SQ_{res}(\mu) = \sum_{i=1}^n \epsilon_j^2 = \sum_{i=1}^n (y_j - \mu)^2$$

$$\frac{dSQ_{res}(\mu)}{d\mu} = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^n y_j}{n} = \bar{y}$$



Estimação

Modelos podem ser escritos como:

Modelo 0 : $Y_j = \mu + \epsilon_j$

parâmetros: μ e σ

Modelo 1 : $Y_{ij} = \mu_i + \epsilon_{ij}$

parâmetros: μ_1, μ_2 e σ

Modelo 2 : $Y_j = \beta_0 + \beta_1 x_j + \epsilon_j$

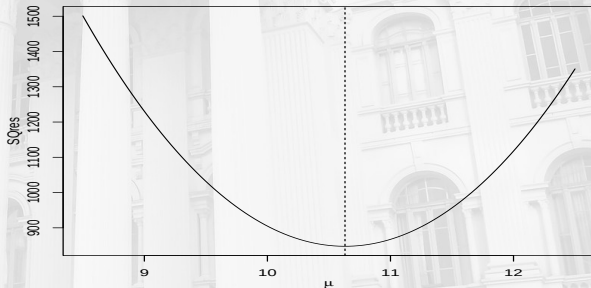
parâmetros: β_0, β_1 e σ

Parâmetros podem ser estimados por mínimos quadrados, por exemplo no Modelo 1:

$$SQ_{res}(\mu) = \sum_{i=1}^n \epsilon_j^2 = \sum_{i=1}^n (y_j - \mu)^2$$

$$\frac{dSQ_{res}(\mu)}{d\mu} = 0$$

$$\hat{\mu} = \frac{\sum_{i=1}^n y_j}{n} = \bar{y}$$



Neste caso, estimadores são os mesmos obtidos pelo método da verossimilhança.

Exemplo cats: expressando incerteza

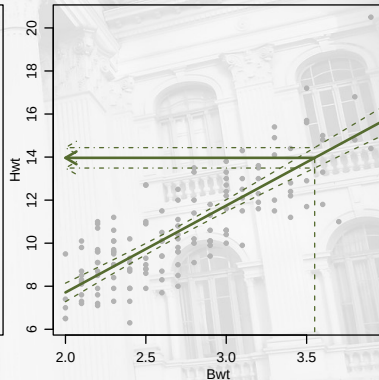
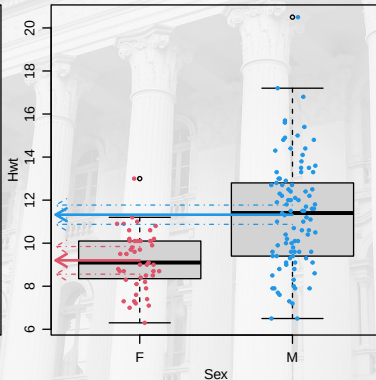
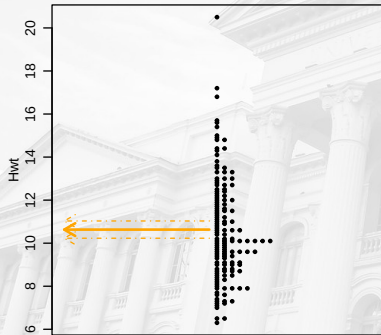
Descritiva, Probabilidades, Inferência, Modelagem

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \mu$$

$$\mu_{ij} = \mu_i, i \in \{F, M\}$$

$$\mu_{ij} = \mu_j = \beta_0 + \beta_1 x_j$$



Exemplo cats: comparando modelos iniciais

	Modelo 0	Modelo 1	Modelo 2
logLik	-332.0	-318.7	-257.1
df	2	3	3
AIC	667.9	643.4	520.1
BIC	673.9	652.3	529.0

Ajustes e penalizações

$$-2\log(Lik) + k \cdot (\#par)$$

usuais: $k = 2(\text{AIC})$ ou $k = \log(n)(\text{BIC})$

Especificação

O modelo de **regressão linear simples** é usualmente escrito da forma:

$$Y_j = \beta_0 + \beta_1 x_j + \epsilon_j$$

$$\epsilon_j \sim N(0, \sigma^2)$$

Equivalentemente, pode ser expresso na forma:

$$Y_j | x_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_j$$

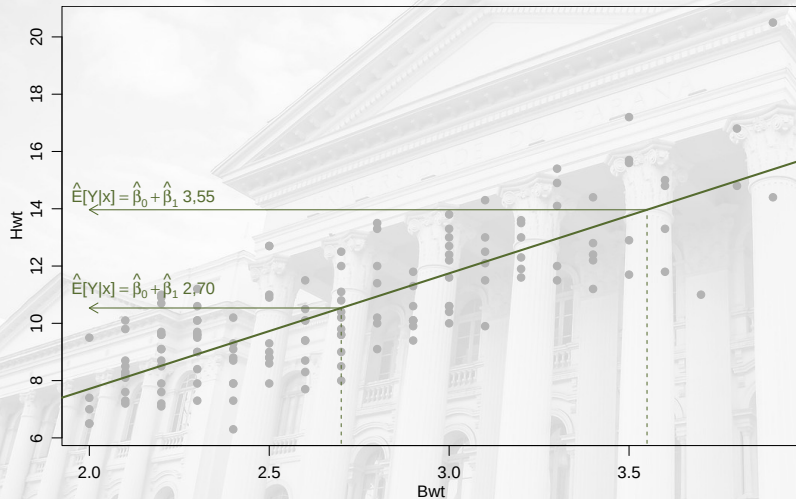
em que

Y_j : observação j correspondente ao valor de x_j

$\mu_j = \beta_0 + \beta_1 x_j \longrightarrow$ para mesmo x_j tem-se a mesma média

$\sigma^2 \longrightarrow$ todas observações possuem a mesma variância

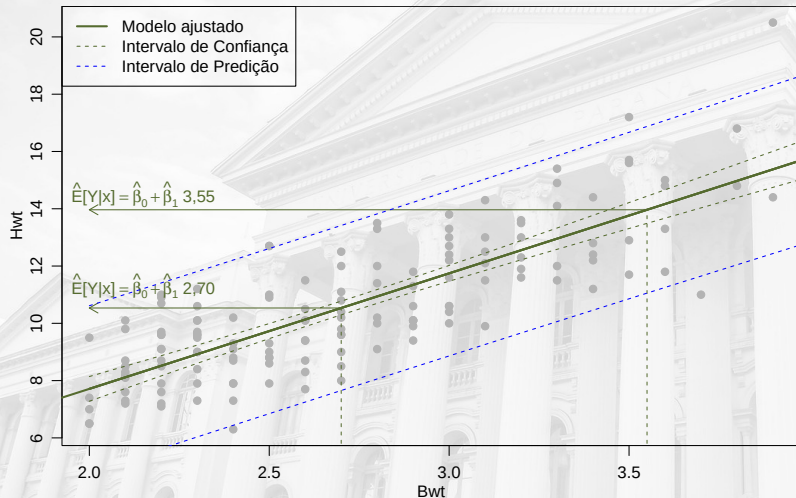
Predição por Regressão Linear Simples



$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_j$$

Regressão Linear Simples



$$Y_j \sim N(\mu_j, \sigma^2)$$
$$\mu_j = \beta_0 + \beta_1 x_j$$

Extendendo o modelo: combinando preditoras

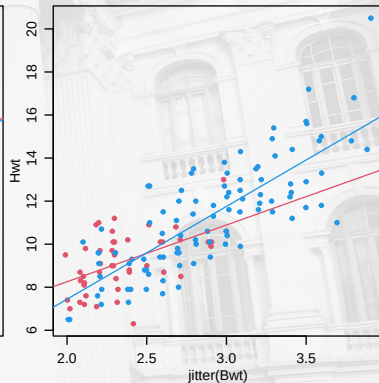
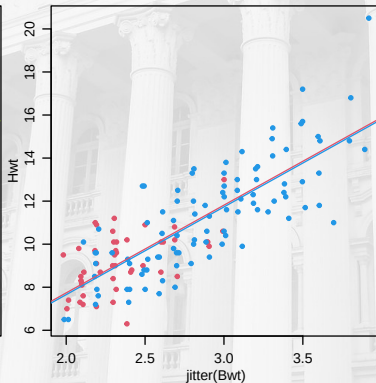
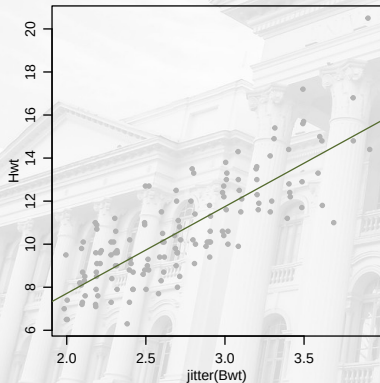
$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \mu_j = \beta_0 + \beta_1 x_{ij}$$

$$\mu_{ij} = \beta_{0i} + \beta_1 x_{ij}$$

$$i \in \{F, M\}$$

$$\mu_{ij} = \beta_{0i} + \beta_{1i} x_{ij}$$



Exemplo **cats**: Modelos ajustados para média e variância.

Mod.	Parâm.	logLik	npar	AIC	BIC	Mod.	Parâm.	logLik	npar	AIC	BIC
0	μ, σ	-332.0	2	667.9	673.9	0.1	μ, σ_i	-327.0	3	660.0	669.0
1	μ_i, σ	-318.7	3	643.4	652.3	1.1	μ_i, σ_i	-308.2	4	624.4	636.3
2	β_0, β_1, σ	-257.1	3	520.1	529.0	2.1	$\beta_0, \beta_1, \sigma_i$	-255.1	4	518.2	530.0
3	$\beta_{0i}, \beta_1, \sigma$	-257.0	4	522.0	533.9	3.1	$\beta_{0i}, \beta_1, \sigma_i$	-255.1	5	520.2	535.0
4	$\beta_{0i}, \beta_{1i}, \sigma$	-255.0	5	520.0	534.8	4.1	$\beta_{0i}, \beta_{1i}, \sigma_i$	-252.3	6	516.6	534.4

Avaliando ajustes sequenciais (aninhados)

Sequência de Modelos 1

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	143	847.63				
Sex	142	705.26	1	142.37	68.48	0.0000
Sex + Bwt	141	299.38	1	405.88	195.24	0.0000
Sex * Bwt	140	291.05	1	8.33	4.01	0.0472

Avaliando ajustes sequenciais (aninhados)

Sequência de Modelos 1

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	143	847.63				
Sex	142	705.26	1	142.37	68.48	0.0000
Sex + Bwt	141	299.38	1	405.88	195.24	0.0000
Sex * Bwt	140	291.05	1	8.33	4.01	0.0472

Sequência de Modelos 2

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	143	847.63				
Bwt	142	299.53	1	548.09	263.64	0.0000
BWt + Sex	141	299.38	1	0.15	0.07	0.7853
Bwt * Sex	140	291.05	1	8.33	4.01	0.0472

Qual sequência faz mais sentido no contexto?

Resumos de um ajuste

Coefficientes

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.9813	1.8428	1.62	0.1080
SexM	-4.1654	2.0618	-2.02	0.0453
Bwt	2.6364	0.7759	3.40	0.0009
SexM:Bwt	1.6763	0.8373	2.00	0.0472

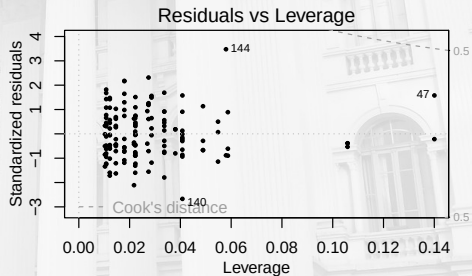
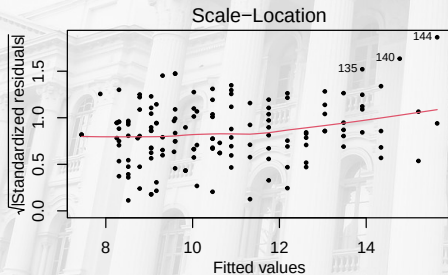
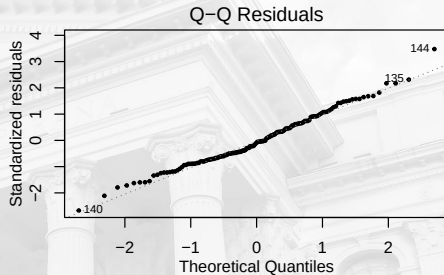
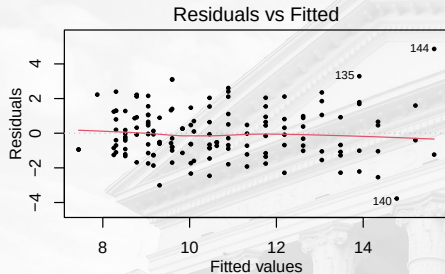
Resumos de um ajuste

Coefficients

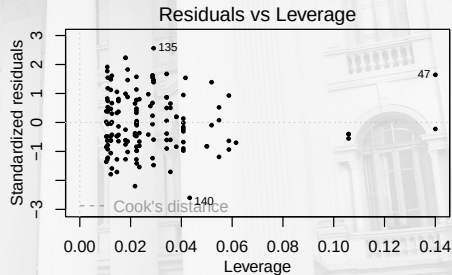
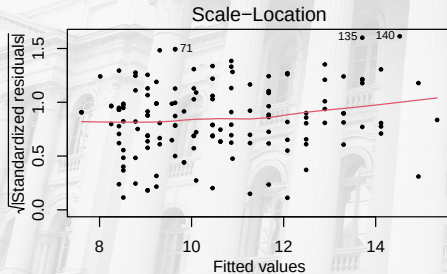
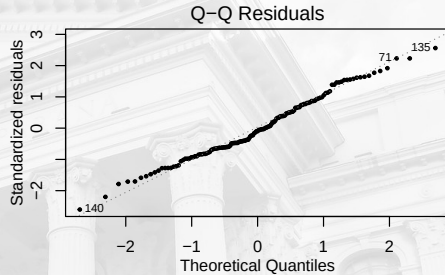
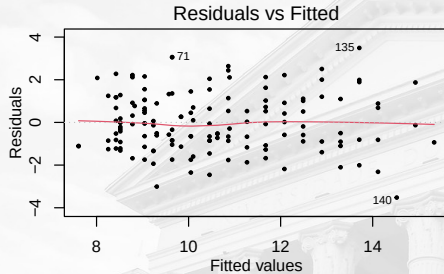
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.9813	1.8428	1.62	0.1080
SexM	-4.1654	2.0618	-2.02	0.0453
Bwt	2.6364	0.7759	3.40	0.0009
SexM:Bwt	1.6763	0.8373	2.00	0.0472

Análise de variância

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sex	1	142.37	142.37	68.48	0.0000
Bwt	1	405.88	405.88	195.24	0.0000
Sex:Bwt	1	8.33	8.33	4.01	0.0472
Residuals	140	291.05	2.08		



Removendo a observação (144)



Removendo uma observação

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	142	749.54				
Sex	141	620.16	1	129.38	67.63	0.0000
Sex + Bwt	140	271.95	1	348.21	182.01	0.0000
Sex * Bwt	139	265.93	1	6.02	3.15	0.0782

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	142	749.54				
Bwt	141	271.95	1	477.58	249.63	0.0000
BWt + Sex	140	271.95	1	0.01	0.00	0.9488
Bwt * Sex	139	265.93	1	6.02	3.15	0.0782

Mudaram conclusões e escolha de modelo?

Removendo uma observação

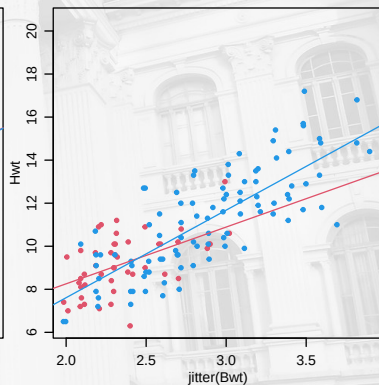
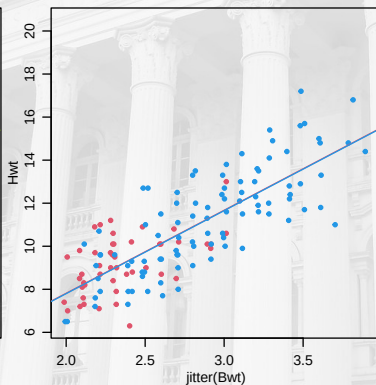
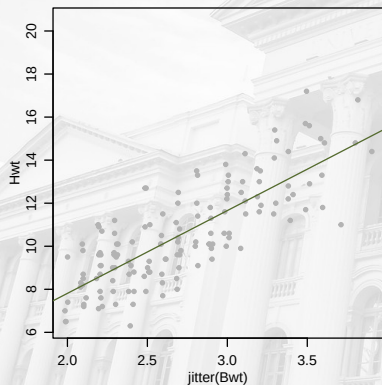
$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \mu_j = \beta_0 + \beta_1 x_{ij}$$

$$\mu_{ij} = \beta_{0i} + \beta_1 x_{ij}$$

$$i \in \{F, M\}$$

$$\mu_{ij} = \beta_{0i} + \beta_{1i} x_{ij}$$



Exemplo **cats**: Modelos (re)ajustados para média e variância.

Mod.	Parâm.	logLik	npar	AIC	BIC	Mod.	Parâm.	logLik	npar	AIC	BIC
0	μ, σ	-321.4	2	646.7	652.6	0.1	μ, σ_i	-317.8	3	641.6	650.5
1	μ_i, σ	-307.8	3	621.6	630.5	1.1	μ_i, σ_i	-299.3	4	606.5	618.4
2	β_0, β_1, σ	-248.9	3	503.7	512.6	2.1	$\beta_0, \beta_1, \sigma_i$	-247.5	4	502.9	514.8
3	$\beta_{0i}, \beta_1, \sigma$	-248.9	4	505.7	517.6	3.1	$\beta_{0i}, \beta_1, \sigma_i$	-247.5	5	504.9	519.7
4	$\beta_{0i}, \beta_{1i}, \sigma$	-247.3	5	504.5	519.3	4.1	$\beta_{0i}, \beta_{1i}, \sigma_i$	-245.4	6	502.8	520.6

Todos os ajustes

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
fitcatso	1	2.00	667.91	673.85	-331.96			
fitcatso.1	2	3.00	660.04	668.95	-327.02	1 vs 2	9.87	0.00
fitcats1	3	3.00	643.43	652.34	-318.72			
fitcats1.1	4	4.00	624.39	636.27	-308.20	3 vs 4	21.04	0.00
fitcats2	5	3.00	520.12	529.03	-257.06	4 vs 5	102.27	0.00
fitcats2.1	6	4.00	518.15	530.03	-255.08	5 vs 6	3.97	0.05
fitcats3	7	4.00	522.05	533.93	-257.02			
fitcats3.1	8	5.00	520.15	535.00	-255.08	7 vs 8	3.90	0.05
fitcats4	9	5.00	519.98	534.83	-254.99			
fitcats4.1	10	6.00	516.60	534.42	-252.30	9 vs 10	5.38	0.02

Todos os (re)ajustes

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
fitcatso	1	2.00	646.71	652.64	-321.36			
fitcatso.1	2	3.00	641.59	650.48	-317.79	1 vs 2	7.12	0.01
fitcats1	3	3.00	621.62	630.50	-307.81			
fitcats1.1	4	4.00	606.53	618.39	-299.27	3 vs 4	17.08	0.00
fitcats2	5	3.00	503.74	512.62	-248.87	4 vs 5	100.80	0.00
fitcats2.1	6	4.00	502.94	514.80	-247.47	5 vs 6	2.79	0.09
fitcats3	7	4.00	505.73	517.58	-248.87			
fitcats3.1	8	5.00	504.93	519.75	-247.47	7 vs 8	2.80	0.09
fitcats4	9	5.00	504.53	519.34	-247.26			
fitcats4.1	10	6.00	502.80	520.58	-245.40	9 vs 10	3.73	0.05

Avaliando ajustes

- Não utilize resultados de forma cega e automática.

Avaliando ajustes

- ▶ Não utilize resultados de forma cega e automática.
- ▶ Considere o contexto e questões de interesse ao construir sequência de modelos.

Avaliando ajustes

- ▶ Não utilize resultados de forma cega e automática.
- ▶ Considere o contexto e questões de interesse ao construir sequência de modelos.
- ▶ Resultados são um guia para discussão.

Avaliando ajustes

- ▶ Não utilize resultados de forma cega e automática.
- ▶ Considere o contexto e questões de interesse ao construir sequência de modelos.
- ▶ Resultados são um guia para discussão.
- ▶ Sex: variável de controle ou de interesse?

Avaliando ajustes

- ▶ Não utilize resultados de forma cega e automática.
- ▶ Considere o contexto e questões de interesse ao construir sequência de modelos.
- ▶ Resultados são um guia para discussão.
- ▶ Sex: variável de controle ou de interesse?
- ▶ Variâncias distintas entre os grupos? Mesmo?

Avaliando ajustes

- ▶ Não utilize resultados de forma cega e automática.
- ▶ Considere o contexto e questões de interesse ao construir sequência de modelos.
- ▶ Resultados são um guia para discussão.
- ▶ Sex: variável de controle ou de interesse?
- ▶ Variâncias distintas entre os grupos? Mesmo?
- ▶ Significância estatística e relevância prática.

Avaliando ajustes

- ▶ Não utilize resultados de forma cega e automática.
- ▶ Considere o contexto e questões de interesse ao construir sequência de modelos.
- ▶ Resultados são um guia para discussão.
- ▶ Sex: variável de controle ou de interesse?
- ▶ Variâncias distintas entre os grupos? Mesmo?
- ▶ Significância estatística e relevância prática.
- ▶ Ponto(s)(dado(s)) extremo(s) deve(m) ser observado(s).

Avaliando ajustes

- ▶ Não utilize resultados de forma cega e automática.
- ▶ Considere o contexto e questões de interesse ao construir sequência de modelos.
- ▶ Resultados são um guia para discussão.
- ▶ Sex: variável de controle ou de interesse?
- ▶ Variâncias distintas entre os grupos? Mesmo?
- ▶ Significância estatística e relevância prática.
- ▶ Ponto(s)(dado(s)) extremo(s) deve(m) ser observado(s).
- ▶ Faça gráficos sempre que possível!

Avaliando ajustes

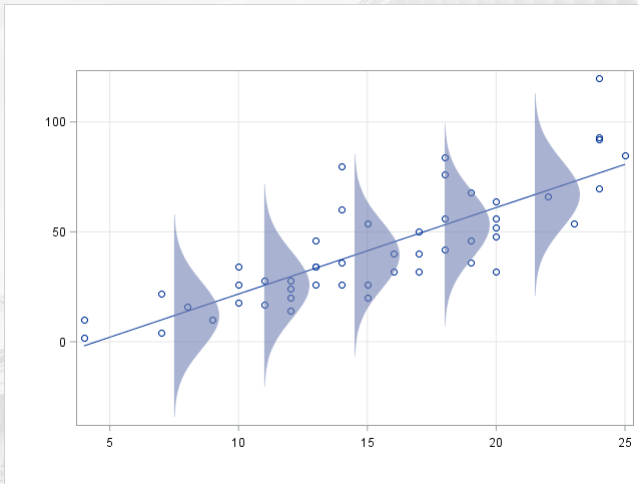
- ▶ Não utilize resultados de forma cega e automática.
- ▶ Considere o contexto e questões de interesse ao construir sequência de modelos.
- ▶ Resultados são um guia para discussão.
- ▶ Sex: variável de controle ou de interesse?
- ▶ Variâncias distintas entre os grupos? Mesmo?
- ▶ Significância estatística e relevância prática.
- ▶ Ponto(s)(dado(s)) extremo(s) deve(m) ser observado(s).
- ▶ Faça gráficos sempre que possível!

Avaliando ajustes

- ▶ Não utilize resultados de forma cega e automática.
- ▶ Considere o contexto e questões de interesse ao construir sequência de modelos.
- ▶ Resultados são um guia para discussão.
- ▶ Sex: variável de controle ou de interesse?
- ▶ Variâncias distintas entre os grupos? Mesmo?
- ▶ Significância estatística e relevância prática.
- ▶ Ponto(s)(dado(s)) extremo(s) deve(m) ser observado(s).
- ▶ Faça gráficos sempre que possível!

Afinal, qual o modelo a ser escolhido?

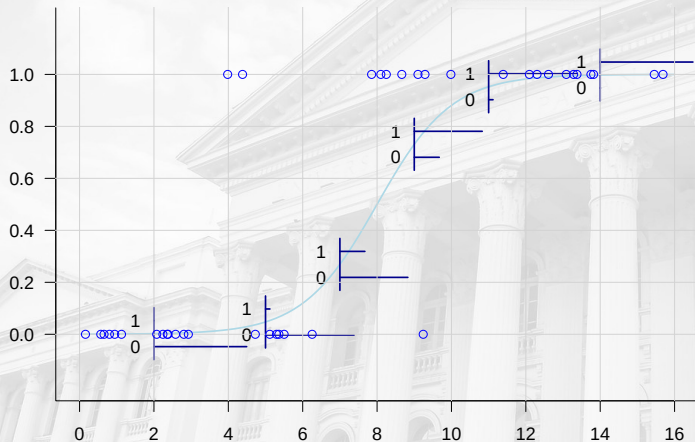
Regressão linear simples



$$Y_j \sim N(\mu_j, \sigma^2)$$
$$\mu_{ij} = \mu_j = \beta_0 + \beta_1 x_j$$

Figura 2. Regressão linear simples.

Regressão para dados binários



$$Y_j \sim \text{Ber}(\theta_j)$$

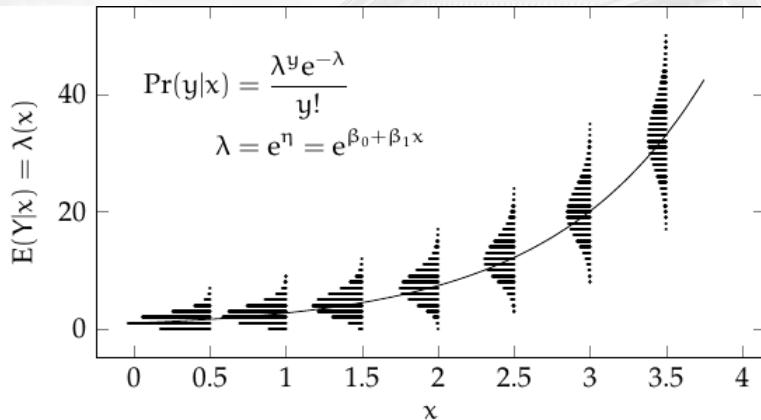
$$g(\theta_j) = \beta_0 + \beta_1 x_j$$

$g(\cdot)$: função de ligação.

$\log(\cdot)$: logaritmo neperiano.

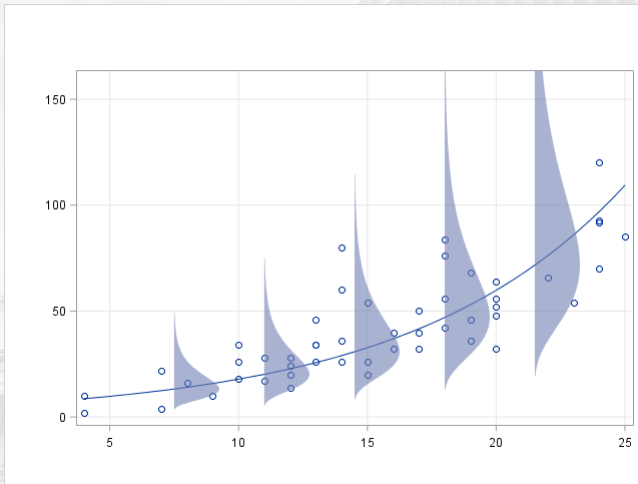
Figura 3. Regressão para dados binários

Regressão de Poisson



$$Y_j \sim \text{P}(\lambda_j)$$
$$\log(\lambda_j) = \beta_0 + \beta_1 x_j$$

Figura 4. Regressão para dados de contagens



$$Y_j \sim \text{G}(\mu_j, \phi)$$
$$\log(\mu_i) = \beta_0 + \beta_1 x_j$$

Figura 5. Regressão para dados com distribuição assimétrica

$$Y_j \sim \text{Distr.}(\mu_j, \phi)$$
$$g(\mu_j) = \eta_j = \beta_0 + \beta_1 x_j$$

- A distribuição de probabilidades da variável resposta é da *família exponencial* (normal, binomial, poisson, gama são "membros").

Modelos lineares generalizados

$$Y_j \sim \text{Distr.}(\mu_j, \phi)$$

$$g(\mu_j) = \eta_j = \beta_0 + \beta_1 x_j$$

- ▶ A **distribuição de probabilidades** da variável resposta é da *família exponencial* (normal, binomial, poisson, gama são "membros").
- ▶ a **função de ligação** relaciona a média da resposta $E[Y|x]$ ao **preditor linear**

$$Y_j \sim \text{Distr.}(\mu_j, \phi)$$
$$g(\mu_j) = \eta_j = \beta_0 + \beta_1 x_j$$

- ▶ A distribuição de probabilidades da variável resposta é da família exponencial (normal, binomial, poisson, gama são "membros").
- ▶ a função de ligação relaciona a média da resposta $E[Y|x]$ ao preditor linear
- ▶ pode haver um parâmetro de dispersão.

Modelos de regressão

A variável resposta Y é explicada pelo preditor x :

$$h(Y_j) \sim \text{Distr.}(\mu(x_j), \phi)$$

$$E[Y|x_j] = \eta_j = f(x_j)$$

- Y_i é a variável resposta (ou uma transformação desta)

Modelos de regressão

A variável resposta Y é explicada pelo preditor x :

$$h(Y_j) \sim \text{Distr.}(\mu(x_j), \phi)$$

$$E[Y|x_j] = \eta_j = f(x_j)$$

- ▶ Y_i é a variável resposta (ou uma transformação desta)
- ▶ para a qual se assume uma distribuição de probabilidades

Modelos de regressão

A variável resposta Y é explicada pelo preditor x :

$$h(Y_j) \sim \text{Distr.}(\mu(x_j), \phi)$$

$$E[Y|x_j] = \eta_j = f(x_j)$$

- ▶ Y_i é a **variável resposta** (ou uma transformação desta)
- ▶ para a qual se assume uma **distribuição de probabilidades**
- ▶ da qual o comportamento resume (como, por exemplo, a **média**)

Modelos de regressão

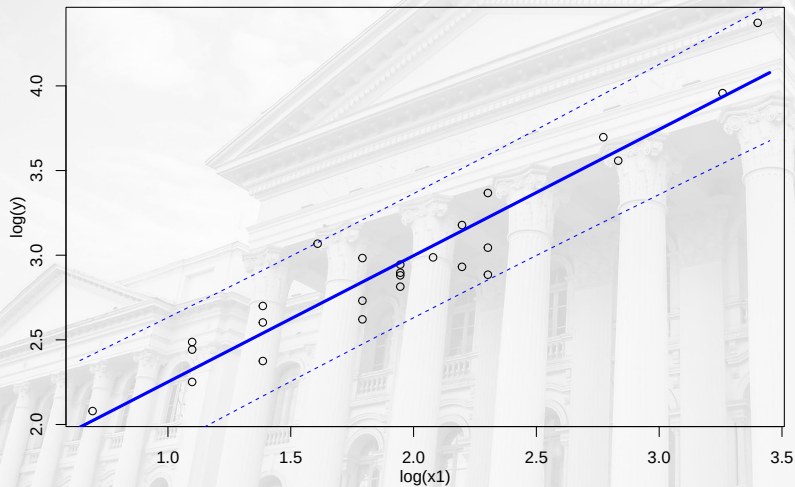
A variável resposta Y é explicada pelo preditor x :

$$h(Y_j) \sim \text{Distr.}(\mu(x_j), \phi)$$

$$E[Y|x_j] = \eta_j = f(x_j)$$

- ▶ Y_i é a **variável resposta** (ou uma transformação desta)
- ▶ para a qual se assume uma **distribuição de probabilidades**
- ▶ da qual o comportamento resume (como, por exemplo, a **média**)
- ▶ é "explicado" por uma função **$f(x_i)$** da preditora x .

Regressões: transformação dos dados



$$\log(Y_j) \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 \log(x_j)$$

Modelos diferentes:

$$Y = \beta_0 + \beta_1 x + \epsilon \quad \log(Y) = \beta_0 + \beta_1 \log(x) + \epsilon$$

$$Y = e^{\beta_0 + \beta_1 \log(x)} \cdot \epsilon^*$$

ou

$$Y \sim N(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 x$$

ou

$$\log(Y) \sim N(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 \log(x)$$

$$Y = e^{\beta_0 + \beta_1 x} + \epsilon$$

ou

$$Y \sim N(\mu, \sigma^2)$$

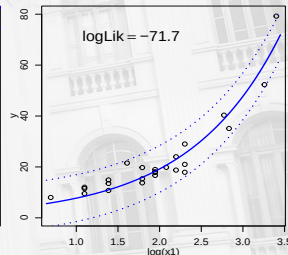
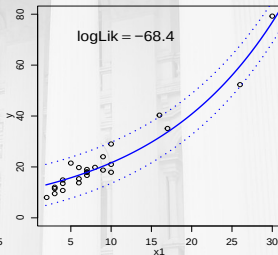
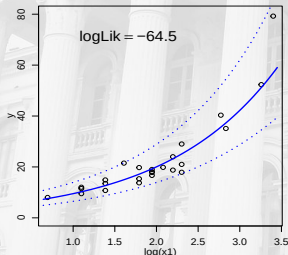
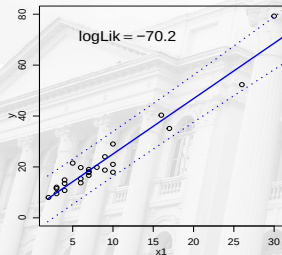
$$\log(\mu) = \beta_0 + \beta_1 x$$

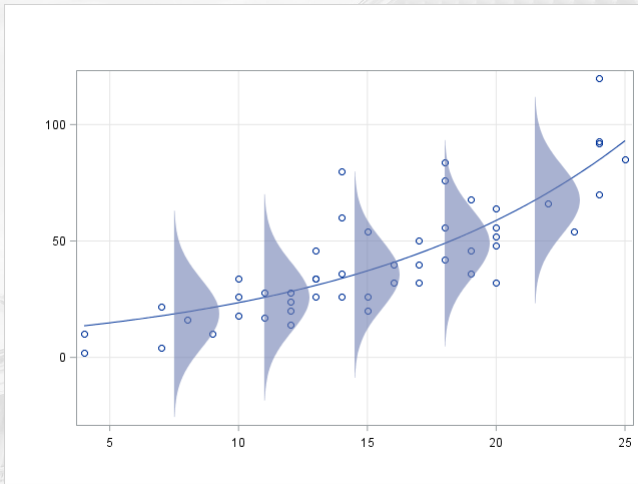
$$Y = e^{\beta_0 + \beta_1 \log(x)} + \epsilon$$

ou

$$Y \sim N(\mu, \sigma^2)$$

$$\log(\mu) = \beta_0 + \beta_1 \log(x)$$





$$Y_j \sim N(\mu_j, \sigma^2)$$

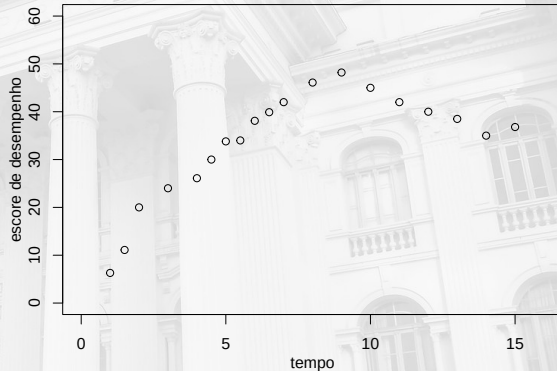
$$\mu_j = \mu_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2$$

Figura 6. Regressão polinomial

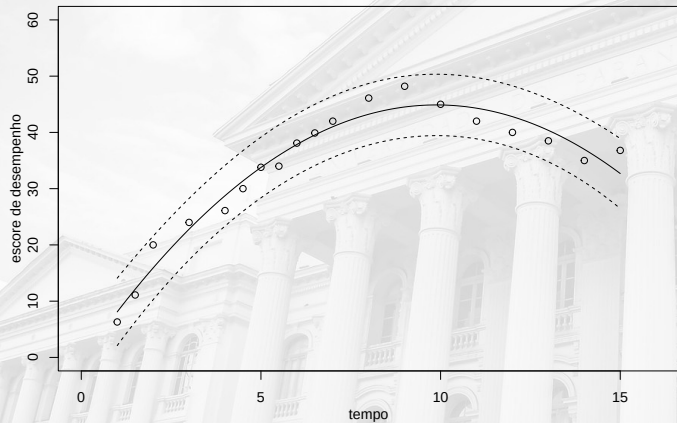
Exemplo - Relação entre tempo de aquecimento e desempenho de ginastas

Tabela 1. Tempo de aquecimento e desempenho

Atleta	t	y	Atleta	t	y
1	1	6.3	11	7	42.0
2	1.5	11.1	12	8	46.1
3	2	20.0	13	9	48.2
4	3	24.0	14	10	45.0
5	4	26.1	15	11	42.0
6	4.5	30.0	16	12	40.0
7	5	33.8	17	13	38.5
8	5.5	34.0	18	14	35
9	6	38.1	19	15	36.8
10	6.5	39.9			



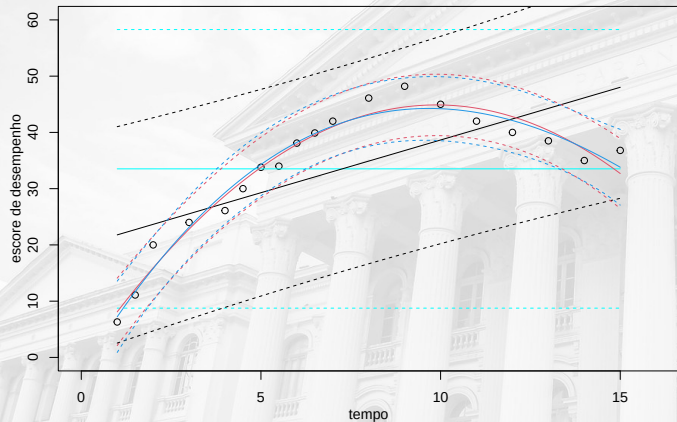
Exemplo - Tempo de aquecimento e desempenho de ginastas



$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

Exemplo - Tempo de aquecimento e desempenho de ginastas



$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \beta_3 x_j^3$$

Exemplo – Tempo de aquecimento e desempenho de ginastas

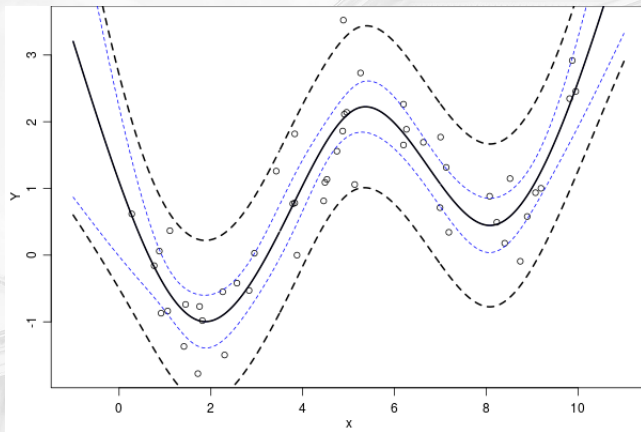
Summary do modelo

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	33.5211	0.5616	59.69	0.0000
poly(tempo, 3)1	34.1835	2.4478	13.96	0.0000
poly(tempo, 3)2	-33.3402	2.4478	-13.62	0.0000
poly(tempo, 3)3	2.3648	2.4478	0.97	0.3493

ANOVA para modelos aninhados

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Modelo0	18	2375.55				
Modelo1	17	1207.04	1	1168.51	195.01	0.0000
Modelo2	16	95.47	1	1111.57	185.51	0.0000
Modelo3	15	89.88	1	5.59	0.93	0.3493

Regressões suavizadas (splines)



$$Y_j \sim N(\mu_j, \sigma^2)$$

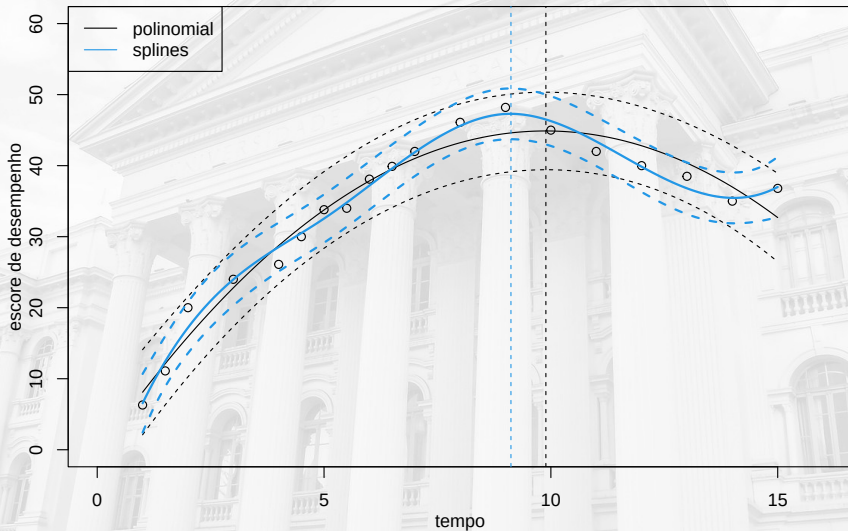
$$\mu_j = s(x_j)$$

Figura 7. Regressão suavizada (I) - splines

Revisitando tempo de aquecimento de ginastas

Ajuste *versus* interpretação

Estimação *versus* predição



Exemplo - Rendimentos de profissionais de certo segmento

Tabela 2. Dados sobre renda, formação e tempo de serviço de profissionais

Profissional	x_1	x_2	y	Profissional	x_1	x_2	y
1	21	0	4554	11	3	1	3414
2	6	0	3195	12	20	1	6928
3	10	0	3539	13	11	1	4651
4	11	0	3742	14	5	1	3836
5	24	0	4707	15	15	1	5595
6	24	0	5034	16	12	1	5172
7	15	0	4331	17	12	1	4732
8	10	0	3748	18	24	1	7612
9	17	0	3887	19	18	1	6478
10	17	0	4436	20	20	1	6779

Exemplo - Renda e tempo de serviço de profissionais

- Dados referentes ao tempo de serviço (x_1 , em anos) e renda (y , em reais) de profissionais de certo segmento.
Outra variável a ser considerada é se os indivíduos possuem ou não curso superior:

$$x_2 = I(x) = \begin{cases} 1, & \text{se possui curso superior} \\ 0, & \text{caso contrário} \end{cases}$$

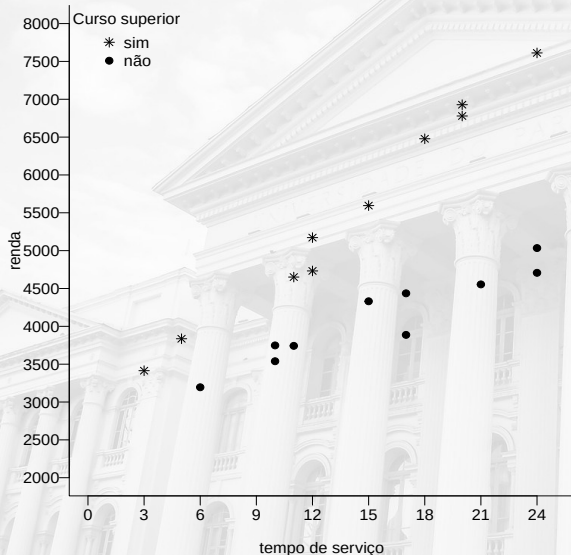
Exemplo - Renda e tempo de serviço de profissionais

- ▶ Dados referentes ao tempo de serviço (x_1 , em anos) e renda (y , em reais) de profissionais de certo segmento.
Outra variável a ser considerada é se os indivíduos possuem ou não curso superior:

$$x_2 = I(x) = \begin{cases} 1, & \text{se possui curso superior} \\ 0, & \text{caso contrário} \end{cases}$$

- ▶ Este exemplo envolve uma variável quantitativa e outra categórica (**análise de covariância**).

Múltiplas Regressões: preditoras categórica e contínua



Objetivos:

- ▶ Há relação entre tempo de serviço e renda?
- ▶ A relação é a mesma para profissionais com e sem curso superior?

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \beta_0 + \beta_1 x_{1j} + \dots$$

$$\dots + \beta_2 x_{2i} + \beta_3 (x_{1j} \cdot x_{2i})$$

- **Modelo 1:** Uma única reta é capaz de explicar a relação entre renda e tempo de serviço para os dois grupos (com e sem ensino superior):

$$y = \beta_0 + \beta_1 x_1 + \epsilon; \quad (1)$$

Modelos

- **Modelo 1:** Uma única reta é capaz de explicar a relação entre renda e tempo de serviço para os dois grupos (com e sem ensino superior):

$$y = \beta_0 + \beta_1 x_1 + \epsilon; \quad (1)$$

- **Modelo 2:** Cada grupo tem sua própria reta, mas as taxas de variação são as mesmas (apenas os interceptos são diferentes):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon. \quad (2)$$

Modelos

- **Modelo 1:** Uma única reta é capaz de explicar a relação entre renda e tempo de serviço para os dois grupos (com e sem ensino superior):

$$y = \beta_0 + \beta_1 x_1 + \epsilon; \quad (1)$$

- **Modelo 2:** Cada grupo tem sua própria reta, mas as taxas de variação são as mesmas (apenas os interceptos são diferentes):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon. \quad (2)$$

- As retas de regressão especificadas pelo modelo 2 são as seguintes:

$$y = \beta_0 + \beta_1 x_1 + \epsilon \text{ (para aqueles que não possuem curso superior);}$$

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon \text{ (para os que possuem curso superior).} \quad (3)$$

- **Modelo 3:** Cada grupo tem sua própria reta, com interceptos e taxas de variação distintos:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon. \quad (4)$$

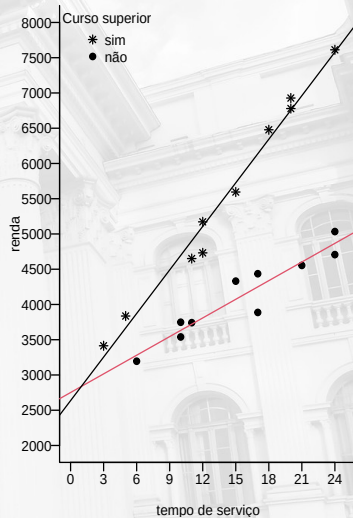
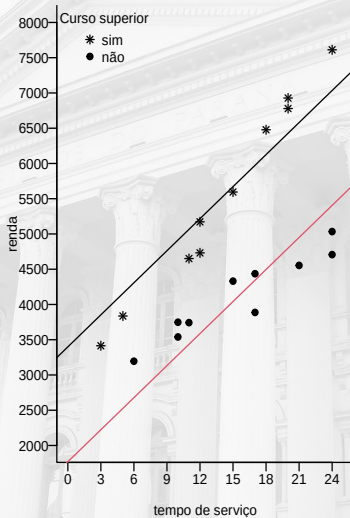
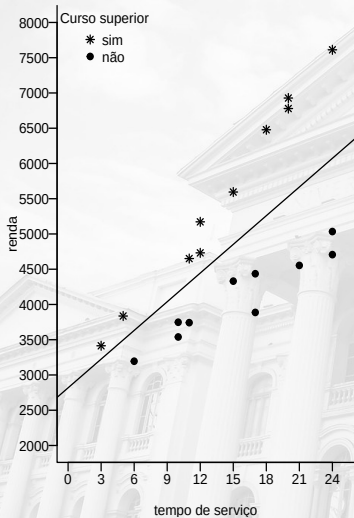
- **Modelo 3:** Cada grupo tem sua própria reta, com interceptos e taxas de variação distintos:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon. \quad (4)$$

- As retas de regressão especificadas pelo modelo 3 são as seguintes:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \epsilon \text{ (para aqueles que não possuem curso superior);} \\ y &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon \text{ (para os que possuem curso superior).} \end{aligned} \quad (5)$$

Modelos ajustados



Escolha de modelos e interpretação

Inferência sobre os β 's respondem aos objetivos:

- ▶ Há relação entre tempo de serviço e renda?
- ▶ A relação é a mesma para profissionais com e sem curso superior?
- ▶ Modelos *encaixados*
 - ▶ Modelo 3:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

- ▶ Modelo 2: $\beta_3 = 0$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

- ▶ Modelo 1: $\beta_3 = 0$ e $\beta_2 = 0$

$$y = \beta_0 + \beta_1 x_1 + \epsilon.$$

Respostas pelos paradigmas de inferência (frequentista, verossimilhança, bayesiano).

Renda: comparação e escolha de modelo

ANOVA para modelos aninhados

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Modelo0	19	30510695.00				
Modelo1	18	16349321.51	1	14161373.49	349.97	0.0000
Modelo2	17	3264224.14	1	13085097.37	323.38	0.0000
Modelo3	16	647426.13	1	2616798.01	64.67	0.0000

Summary do modelo escolhido

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2751.0668	178.2765	15.43	0.0000
x1	88.1441	10.7446	8.20	0.0000
x2	-114.7491	235.0906	-0.49	0.6321
x1:x2	117.8118	14.6500	8.04	0.0000

Renda: coeficientes e modelos ajustados

Expressões do modelo ajustado:

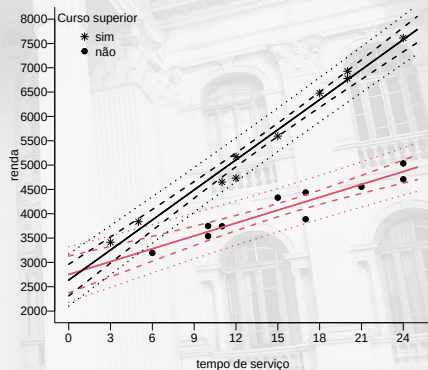
$y = \beta_0 + \beta_1 x_1 + \epsilon$ (para aqueles que não possuem curso superior);

$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1 + \epsilon$ (para os que possuem curso superior).

beta0	beta1	beta2	beta3
2751.07	88.14	-114.75	117.81

Interpretação do modelo ajustado:

- ▶ Sem curso superior
 - ▶ renda inicial: $\hat{\beta}_0 = 2751.07$
 - ▶ incremento anual de renda: $\hat{\beta}_1 = 88.14$
- ▶ Com curso superior
 - ▶ renda inicial: $\hat{\beta}_0 + \hat{\beta}_2 = 2636.32$
 - ▶ incremento anual de renda: $\hat{\beta}_1 + \hat{\beta}_3 = 205.96$



Regressão múltipla

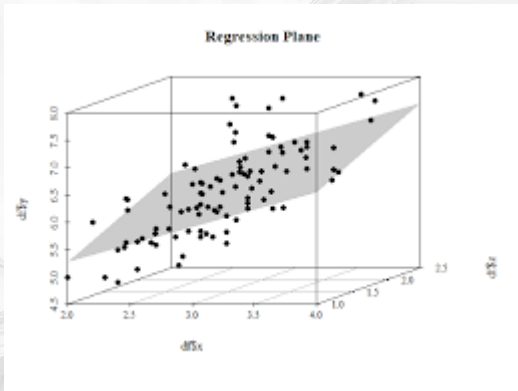


Figura 8. Regressão múltipla

$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots$$

Exemplo - Predição de massa gorda

- ▶ Base de dados fat do pacote faraway. O objetivo é ajustar um modelo de regressão para prever a porcentagem de massa gorda de um indivíduo (brozek) com base nas seguintes variáveis:
- ▶ neck: Circunferência do pescoço;
- ▶ chest: Circunferência do peito;
- ▶ abdom: Circunferência do abdomen;
- ▶ hip: Circunferência do quadril;
- ▶ thigh: Circunferência da coxa;
- ▶ knee: Circunferência do joelho;
- ▶ ankle: Circunferência do tornozelo;
- ▶ biceps: Circunferência do bíceps estendido;
- ▶ forearm: Circunferência do antebraço;
- ▶ wrist: Circunferência do pulso.

Exemplo - Predição de massa gorda

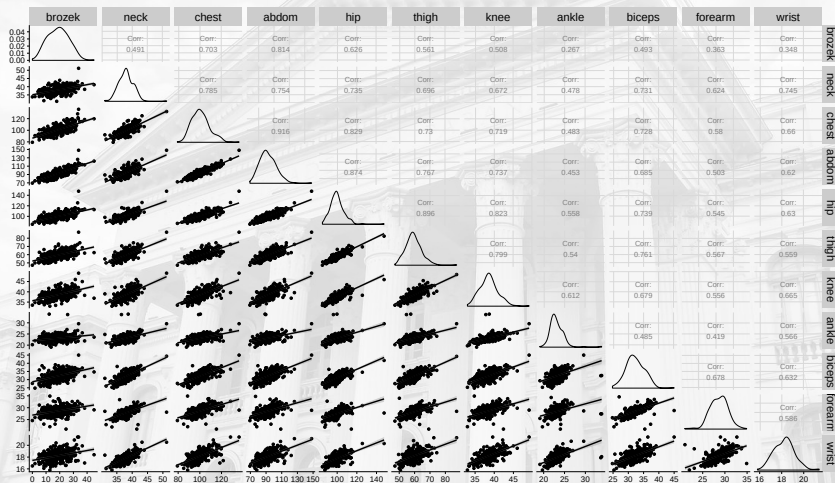


Figura 9. Análise descritiva.

Exemplo - Predição de massa gorda

$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_{10} x_{10j}$$

- De maneira geral, as medidas realizadas estão (positivamente) correlacionadas com o percentual de massa gorda;

Exemplo - Predição de massa gorda

$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_{10} x_{10j}$$

- ▶ De maneira geral, as medidas realizadas estão (positivamente) correlacionadas com o percentual de massa gorda;
- ▶ O ajuste de um modelo de regressão linear múltipla permitiria prever o percentual de massa gorda com base nas demais medidas;

Exemplo - Predição de massa gorda

$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_{10} x_{10j}$$

- ▶ De maneira geral, as medidas realizadas estão (positivamente) correlacionadas com o percentual de massa gorda;
- ▶ O ajuste de um modelo de regressão linear múltipla permitiria prever o percentual de massa gorda com base nas demais medidas;
- ▶ Além disso, podemos notar, conforme esperado, que as medidas de circunferências estão correlacionadas entre si;

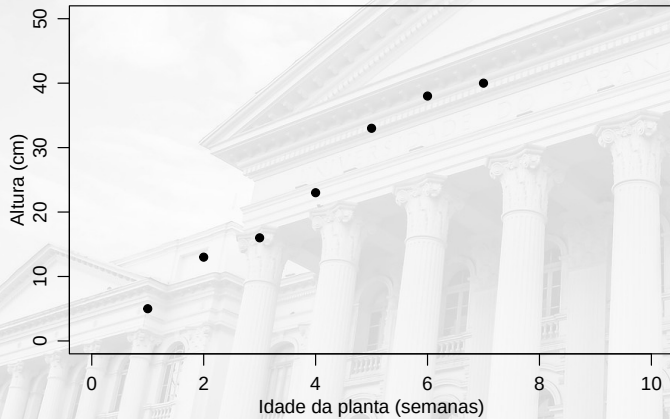
Exemplo - Predição de massa gorda

$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_{10} x_{10j}$$

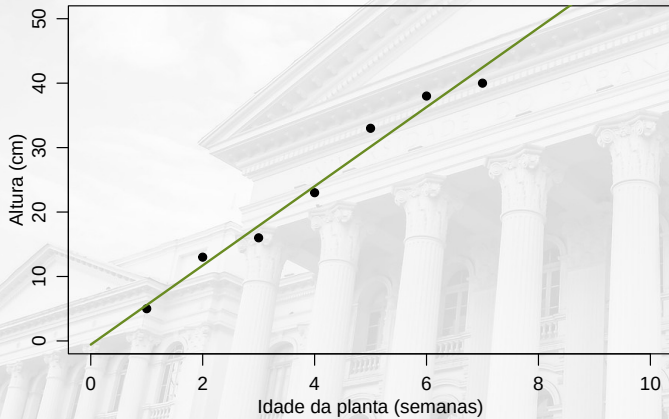
- ▶ De maneira geral, as medidas realizadas estão (positivamente) correlacionadas com o percentual de massa gorda;
- ▶ O ajuste de um modelo de regressão linear múltipla permitiria prever o percentual de massa gorda com base nas demais medidas;
- ▶ Além disso, podemos notar, conforme esperado, que as medidas de circunferências estão correlacionadas entre si;
- ▶ Além disso, a análise de regressão permitiria avaliar a associação de cada medida com o percentual de massa gorda (ajustada pelos valores das demais medidas), e, eventualmente, selecionar as mais importantes para predição.

Regressão Linear ou não linear?



$$Y_j \sim N(\mu_j, \sigma^2)$$
$$\mu_j = ???$$

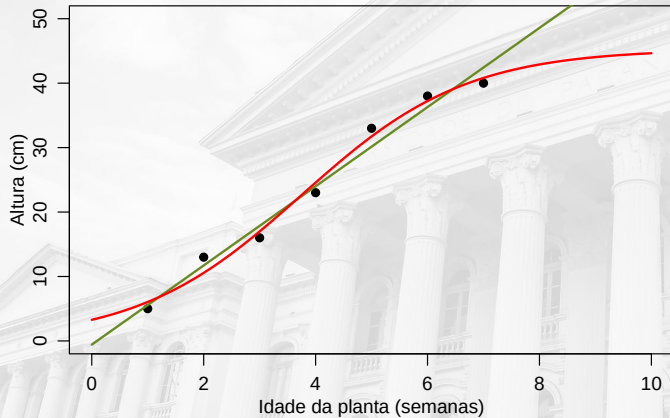
Regressão Linear ou não linear?



$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_j$$

Regressão Linear ou não linear?



$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \beta_0 + \beta_1 x_j$$

$$\mu_j = f(x_j)$$

Um modelo não linear

Modelo de crescimento logístico:

$$f(x) = \frac{A}{1 + \exp(-(x - C)/S)}$$

Um modelo não linear

Modelo de crescimento logístico:

$$f(x) = \frac{A}{1 + \exp(-(x - C)/S)}$$

Reescrevendo com letras **romanas para observadas** e **gregas para estimadas**:

$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \frac{\beta_0}{1 + \exp(-(x_j - \beta_1)/\beta_2)}$$

Um modelo não linear

Modelo de crescimento logístico:

$$f(x) = \frac{A}{1 + \exp(-(x - C)/S)}$$

Reescrevendo com letras **romanas para observadas** e **gregas para estimadas**:

$$Y_j \sim N(\mu_j, \sigma^2)$$

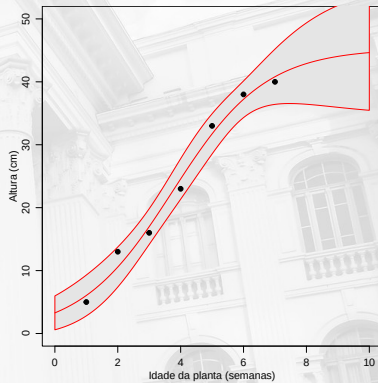
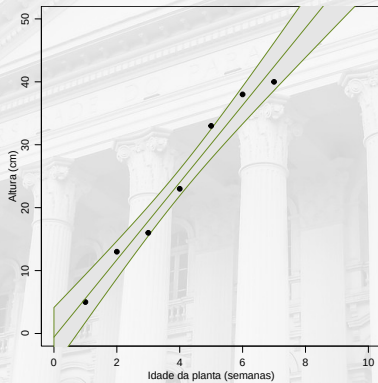
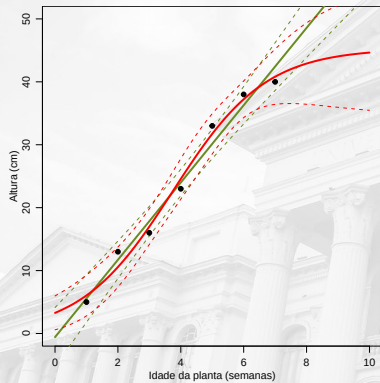
$$\mu_j = \frac{\beta_0}{1 + \exp(-(x_j - \beta_1)/\beta_2)}$$

Interpretações dos parâmetros:

- ▶ $A(\beta_0)$: assíntota: *até onde a planta cresce*
- ▶ $C(\beta_1)$: tempo de meia vida: *tempo para atingir metade do tamanho máximo*
- ▶ $S(\beta_2)$: taxa: *velocidade que a planta cresce*

Ajuste de modelos não lineares por métodos numéricos que exigem "chutes" iniciais.
As interpretações dos parâmetros podem ajudar na definição dos valores iniciais.

Linear e não linear



- Um modelo de regressão é linear se

$$\frac{\partial f(\mathbf{x}; \boldsymbol{\beta})}{\partial \beta_j} \quad (6)$$

não depender de β_j , para todo $j = 0, 1, \dots, p$.

- Ou seja, num modelo de regressão linear, ao se derivar parcialmente em relação a um parâmetro β_j , o resultado não depende de qualquer dos $\beta_{j'}$.

Modelos de regressão linear podem ser escritos na forma matricial:

$$\underset{n \times 1}{\mathbf{Y}} = \underset{n \times p}{\mathbf{X}} \cdot \underset{p \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\boldsymbol{\epsilon}}$$

Lineares e Não Lineares

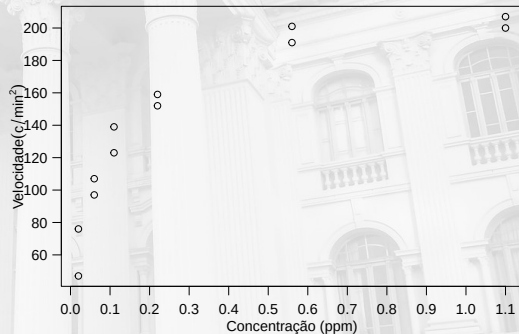
1. $y = \beta_0 + \beta_1 x + \epsilon$
2. $y = \beta_0 + \beta_1 \ln(x) + \epsilon$
3. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$
4. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
5. $y = \beta_0 + \beta_1 \ln(x_1) + \frac{\beta_2}{x_2} + \beta_3 x_1 x_2 + \epsilon$
6. $y = \beta_0 + \beta_1 e^{\beta_2 x_2} + \epsilon$
7. $y = \frac{\beta_1}{1 + e^{\beta_2 x}} + \epsilon$
8. $y = \beta_0 \sin(\beta_1 + \beta_2 x) + \epsilon$

Exemplo - Reação química

Velocidade de uma reação enzimática (y , expressa em *contagem/min*²) e a concentração de certo substrato (x , em *partes por milhão*) em 12 repetições de um experimento.

Tabela 3. Dados - reação química

Repetição	x	y	Repetição	x	y
1	0.02	47	7	0.02	76
2	0.06	97	8	0.06	107
3	0.11	123	9	0.11	139
4	0.22	152	10	0.22	159
5	0.56	191	11	0.56	201
6	1.10	200	12	1.10	207



Exemplo - Reação química

- Observamos uma relação não linear entre as variáveis.



Exemplo - Reação química

- ▶ Observamos uma relação não linear entre as variáveis.
- ▶ Ao invés de considerar um modelo polinomial, podemos recorrer à teoria da química e utilizar o **modelo de Michaelis-Menten**, que descreve adequadamente problemas de cinética enzimática.

Exemplo - Reação química

- ▶ Observamos uma relação não linear entre as variáveis.
- ▶ Ao invés de considerar um modelo polinomial, podemos recorrer à teoria da química e utilizar o **modelo de Michaelis-Menten**, que descreve adequadamente problemas de cinética enzimática.
- ▶ Vamos considerar $f(x, \beta) = \frac{\beta_1 x}{x + \beta_2}$, ajustando o seguinte modelo de regressão não linear:

Exemplo - Reação química

- ▶ Observamos uma relação não linear entre as variáveis.
- ▶ Ao invés de considerar um modelo polinomial, podemos recorrer à teoria da química e utilizar o **modelo de Michaelis-Menten**, que descreve adequadamente problemas de cinética enzimática.
- ▶ Vamos considerar $f(x, \beta) = \frac{\beta_1 x}{x + \beta_2}$, ajustando o seguinte modelo de regressão não linear:

$$y = f(x, \beta) + \epsilon$$

ou

$$y = \frac{\beta_1 x}{x + \beta_2} + \epsilon$$

$$Y_j \sim N(\mu_j, \sigma^2)$$

$$\mu_j = \frac{\beta_1 x}{x + \beta_2}$$

Ajuste do modelo

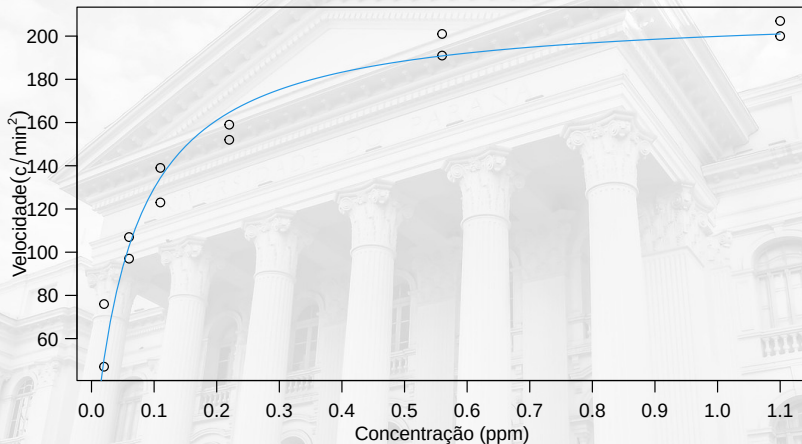


Figura 10. Dados de reação química e modelo não linear ajustado

Exemplo - Reação química

```
rquim.fit <- nls(y ~ b1*x/(x+b2), data = rquim, start = list(b1=200, b2=0.2))
summary(rquim.fit)
```

```
##
## Formula: y ~ b1 * x/(x + b2)
##
## Parameters:
##      Estimate Std. Error t value Pr(>|t|)
## b1 2.127e+02  6.947e+00  30.615 3.24e-11
## b2 6.412e-02  8.281e-03   7.743 1.57e-05
##
## Residual standard error: 10.93 on 10 degrees of freedom
##
## Number of iterations to convergence: 5
## Achieved convergence tolerance: 5.563e-06
```

	Estimate	Std. Error	t value	Pr(> t)
b1	212.68	6.95	30.61	0.00
b2	0.06	0.01	7.74	0.00

Exemplo - Reação química

Seria o mesmo modelo?

Modelo não linear:

$$y = \frac{\beta_1 x}{x + \beta_2} + \epsilon.$$

Exemplo - Reação química

Seria o mesmo modelo?

Modelo não linear:

$$y = \frac{\beta_1 x}{x + \beta_2} + \epsilon.$$

E reexpressando o preditor:

$$\frac{1}{y} = \frac{x + \beta_2}{\beta_1 x} + \epsilon^*$$

$$\frac{1}{y} = \frac{x}{\beta_1 x} + \frac{\beta_2}{\beta_1 x} + \epsilon^*$$

$$y^* = \beta_0^* + \beta_1^* x^* + \epsilon^*$$

em que:

$$y^* = \frac{1}{y} ; \quad x^* = \frac{1}{x} ; \quad \beta_0^* = \frac{1}{\beta_1} ; \quad \beta_1^* = \frac{\beta_2}{\beta_1}$$

Modelo de regressão linear simples com transformação e reparametrização.

Ajustes dos modelos não linear e linearizado

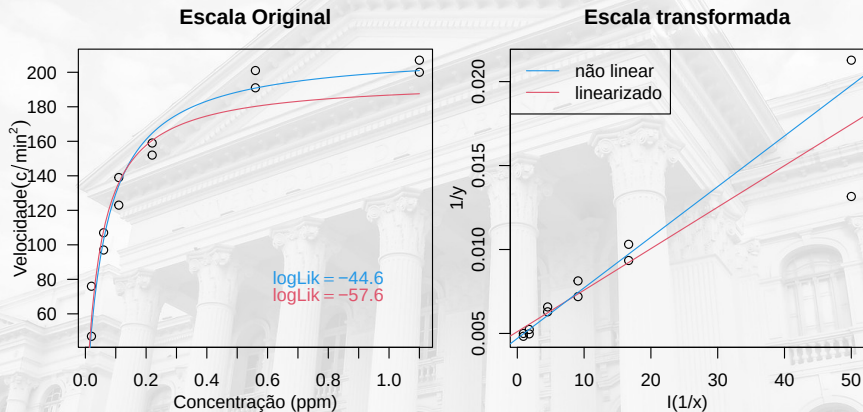
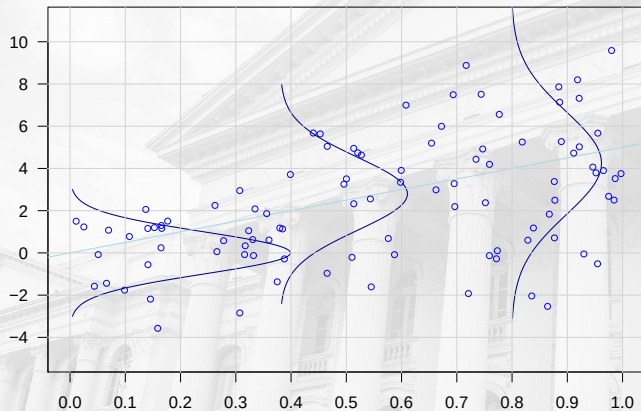


Figura 11. Dados de reação química e ajustes dos modelos

Modelo não-linear (base teórica) ajustado na **escala original** e linear (empírico) ajustado na **escala transformada**.

Regressão Heterocedástica



$$Y_j \sim N(\mu_j, \sigma_j^2)$$

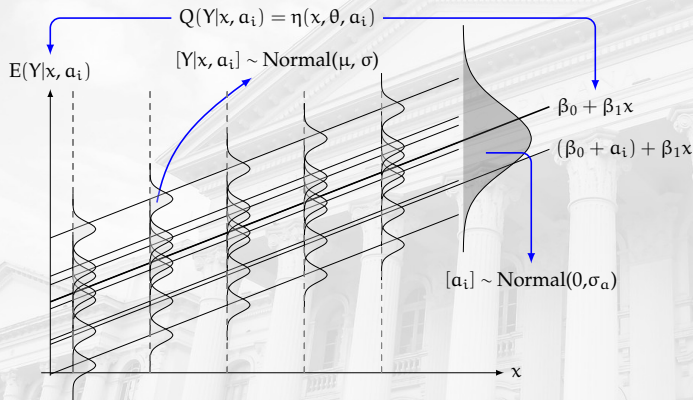
$$\mu_j = \beta_0 + \beta_1 x_j$$

$$\sigma_j^2 = \gamma z_j$$

(caso particular : $z_i = x_i$)

Figura 12. Regressão com erros heterocedásticos (variância não constante)

Regressões: parâmetros aleatórios (variáveis latentes)



$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = (\beta_0 + \textcolor{red}{a}_i) + \beta_1 x_j$$

$$\textcolor{red}{a}_i \sim N(0, \sigma_a^2)$$

Figura 13. Regressão com efeito aleatório

Regressão: especificação geral

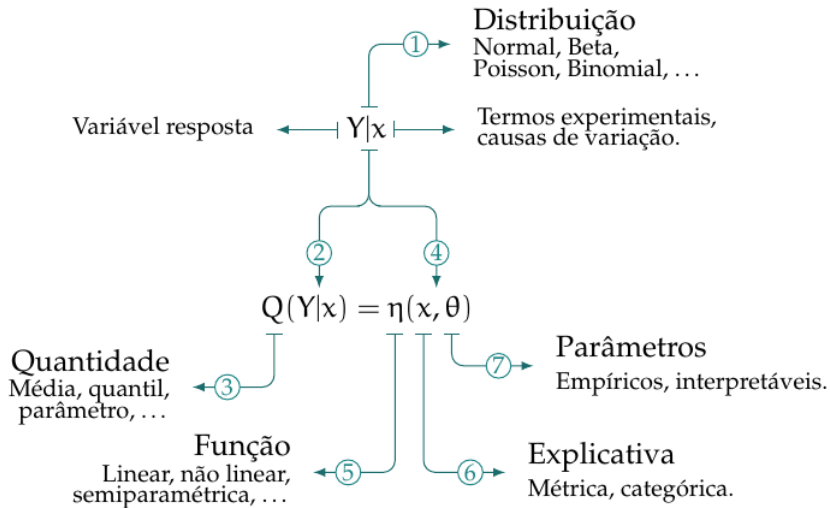


Figura 14. Modelo de regressão de forma geral.

Generalizando ainda mais o modelo de regressão

- A variável resposta pode ter um valor intervalar (ou *censurado*).
Exemplo: análise de sobrevivência (tema de outra aula).

Generalizando ainda mais o modelo de regressão

- ▶ A variável resposta pode ter um valor intervalar (ou *censurado*).
Exemplo: análise de sobrevivência (tema de outra aula).
- ▶ Os preditores podem conter não apenas efeitos (variáveis) fixas (ou fixadas) mas também incluir termos aleatórios (não observados).

Generalizando ainda mais o modelo de regressão

- ▶ A variável resposta pode ter um valor intervalar (ou *censurado*).
Exemplo: análise de sobrevivência (tema de outra aula).
- ▶ Os preditores podem conter não apenas efeitos (variáveis) fixas (ou fixadas) mas também incluir termos aleatórios (não observados).
- ▶ As observações podem ser não independentes (dados longitudinais, espaciais, agrupados, dentre outras possíveis estruturas de dependência).

Generalizando ainda mais o modelo de regressão

- ▶ A variável resposta pode ter um valor intervalar (ou *censurado*). Exemplo: análise de sobrevivência (tema de outra aula).
- ▶ Os preditores podem conter não apenas efeitos (variáveis) fixas (ou fixadas) mas também incluir termos aleatórios (não observados).
- ▶ As observações podem ser não independentes (dados longitudinais, espaciais, agrupados, dentre outras possíveis estruturas de dependência).
- ▶ O modelo usual relaciona a média a covariáveis. Pode se relacionar a variância (reg. heterocedástica) ou também outros parâmetros ou características da distribuição da variável respostas (GAMLSS).

Generalizando ainda mais o modelo de regressão

- ▶ A variável resposta pode ter um valor intervalar (ou *censurado*). Exemplo: análise de sobrevivência (tema de outra aula).
- ▶ Os preditores podem conter não apenas efeitos (variáveis) fixas (ou fixadas) mas também incluir termos aleatórios (não observados).
- ▶ As observações podem ser não independentes (dados longitudinais, espaciais, agrupados, dentre outras possíveis estruturas de dependência).
- ▶ O modelo usual relaciona a média a covariáveis. Pode se relacionar a variância (reg. heterocedástica) ou também outros parâmetros ou características da distribuição da variável respostas (GAMLSS).
- ▶ Várias outras possíveis extensões.

All models are wrong but some are useful

George Box

No matter how beautiful your theory, no matter how clever you are or what your name is, if it disagrees with experiment, it's wrong.

Richard Feynman

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

John W. Tukey