

ESTATÍSTICA PARA SAÚDE COLETIVA

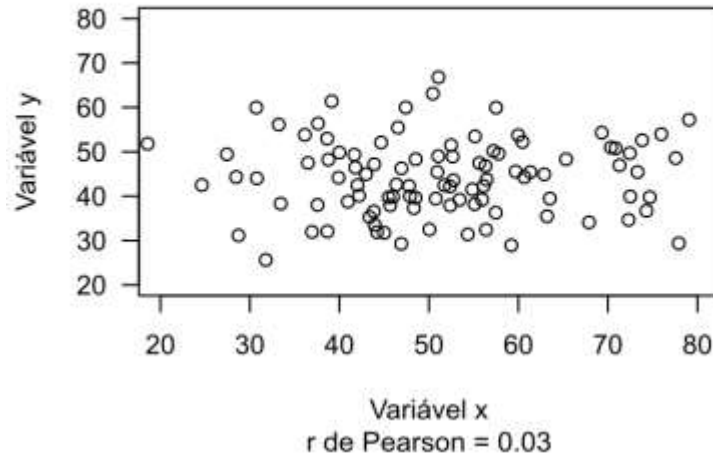
Aula 15

Revisão + feedback lista 13 e 14

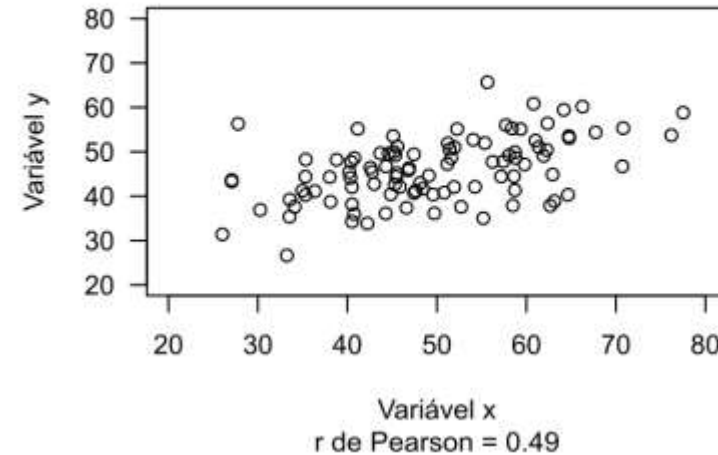
Correlações

Compare diferentes valores de correlações

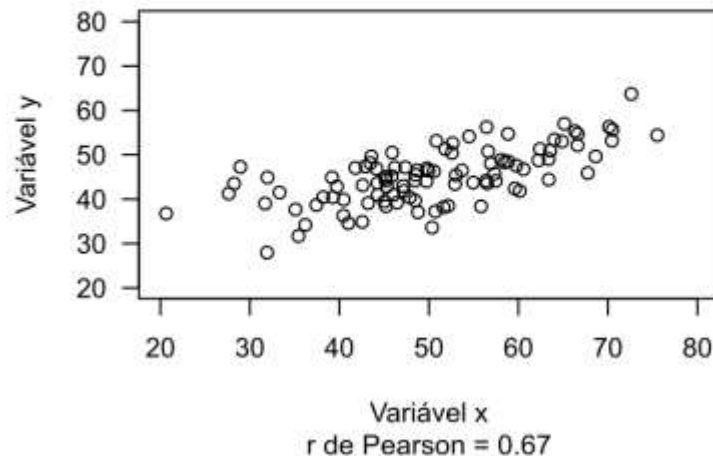
Correlação pequena



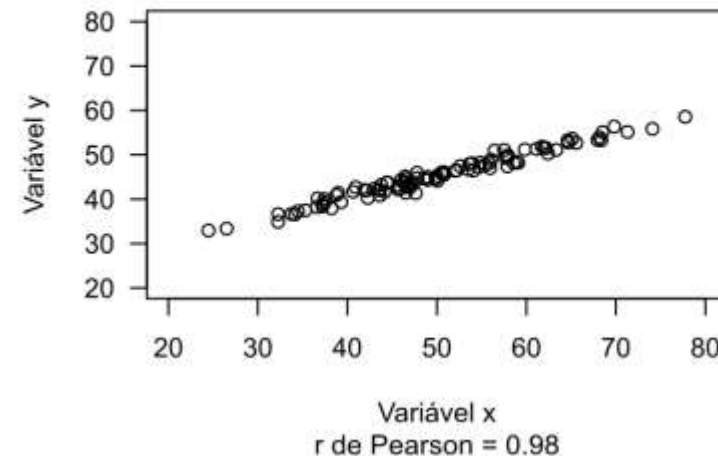
Correlação fraca



Correlação moderada



Correlação forte





Medida de correlação

- Coeficiente de correlação de Pearson (só usado quando você tem duas variáveis quantitativas)
- Pode assumir qualquer valor entre -1 e +1
- Interpretação do valor observado, segundo Vieira (2008)
 - De 0 até 0,25: correlação pequena
 - De 0,25 até 0,5: correlação fraca
 - De 0,5 até 0,75: correlação moderada
 - Acima de 0,75: correlação forte
- Obs: essa classificação é subjetiva e outros autores assumem outros níveis para considerar as intensidades das relações

Lista 13 pergunta 6

Pergunta 6. Assumindo um nível de significância de 5%, existe correlação entre o número de filhos (Coluna NumeroFilhos) e idade dos pais (Coluna Idade)?

1  pontos

- ☐ Não há associação entre as variáveis
- ☒ Há (SIM) associação entre as variáveis 
- ☐ Não consegui responder essa pergunta porque não entendi como aplicar o teste.
- ☐ Não consegui responder essa pergunta porque não tenho computador disponível para executar o teste.
- ☐ Não sei/Não quero responder essa pergunta.

Script

```
# Exercício 6
```

```
# Pearson
```

```
cor.test(tabela1$NumeroFilhos,tabela1$Idade)
```

```
plot(tabela1$NumeroFilhos,tabela1$Idade)
```

```
# Kendall
```

```
cor.test(tabela1$NumeroFilhos,tabela1$Idade,method="kendall")
```

```
boxplot(tabela1$Idade~tabela1$NumeroFilhos)
```

Teste por correlação de Pearson

Pearson's product-moment correlation

data: tabela1\$NumeroFilhos and tabela1\$Idade

t = 11.3, df = 98, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.6521060 0.8264914

sample estimates:

cor
0.7521807

Resultado: P menor que 0.05

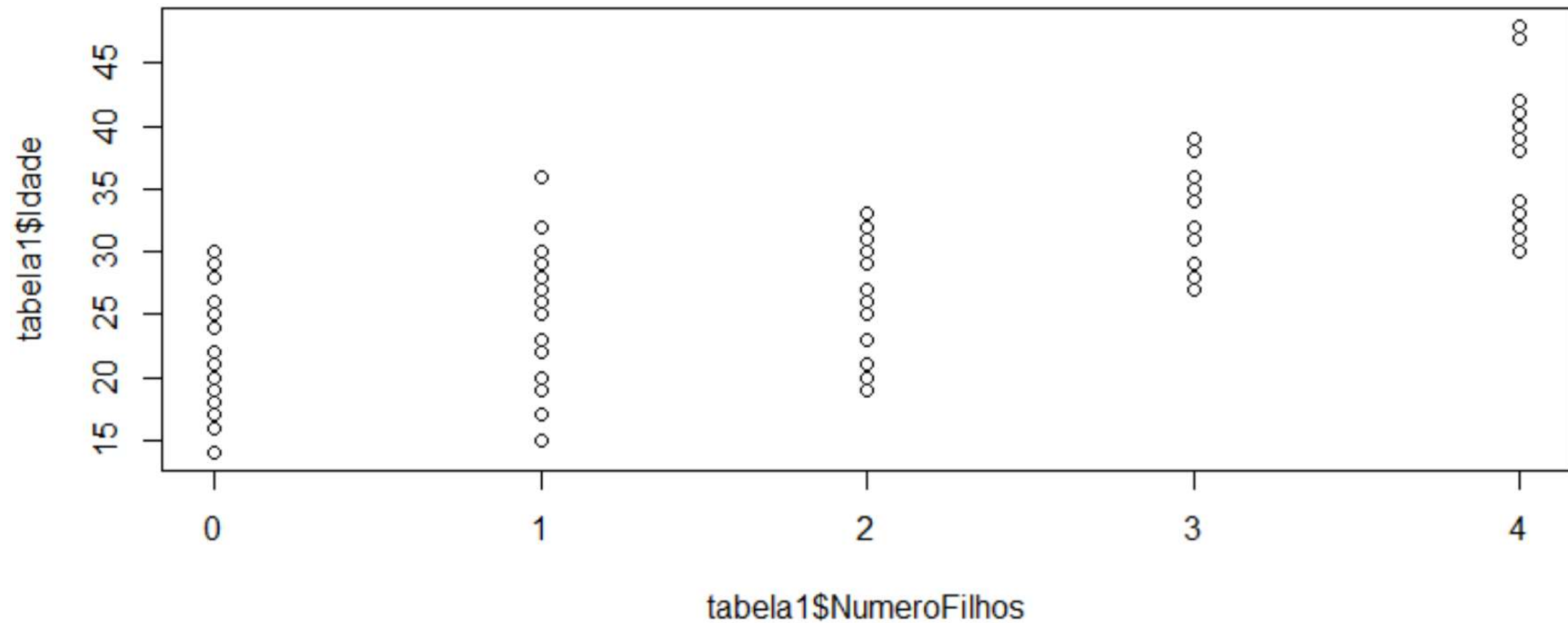
Conclusão: H_1 é verdadeira

Hipóteses testadas

H_0 : Não há correlação entre as variáveis

H_1 : Há (SIM) correlação entre as variáveis

Teste por correlação de Pearson



Teste por correlação de Kendall

Kendall's rank correlation tau

data: tabela1\$NumeroFilhos and tabela1\$Idade

$z = 8.1122$, $p\text{-value} = 4.972e-16$

alternative hypothesis: true tau is not equal to 0
sample estimates:

tau
0.6103924

Resultado: P menor que 0.05

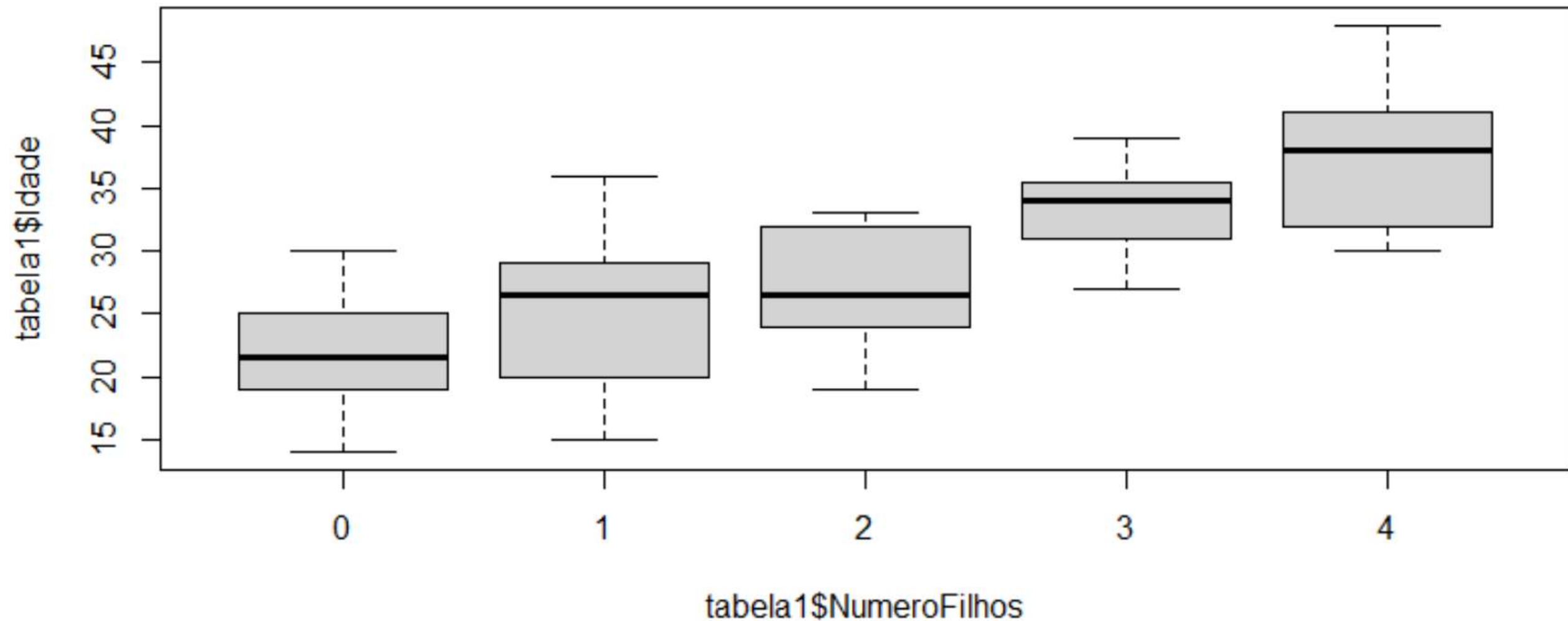
Conclusão: H_1 é verdadeira

Hipóteses testadas

H_0 : Não há correlação entre as variáveis

H_1 : Há (SIM) correlação entre as variáveis

Teste por correlação de Pearson



Lista 13 pergunta 7

Pergunta 7. Assumindo um nível de significância de 5%, existe correlação entre o número de filhos (Coluna NumeroFilhos) e classe social das famílias (Coluna ClasseSocial)?

1  pontos

☒ Não há associação entre as variáveis



☐ Há (SIM) associação entre as variáveis

☐ Não consegui responder essa pergunta porque não entendi como aplicar o teste.

☐ Não consegui responder essa pergunta porque não tenho computador disponível para executar o teste.

☐ Não sei/Não quero responder essa pergunta.

Script

```
# Exercício 7
```

```
cor.test(rank(c(tabela1$ClasseSocial)),tabela1$NumeroFilhos,method="kendall")
```

```
boxplot(tabela1$NumeroFilhos~tabela1$ClasseSocial,ylab="Numero de  
filhos",xlab="Classe social", las=1)
```

Teste por correlação de Kendall

```
> cor.test(rank(c(tabela1$ClasseSocial)), tabela1$NumeroFilhos, method="kendall")
```

Kendall's rank correlation tau

data: rank(c(tabela1\$ClasseSocial)) and tabela1\$NumeroFilhos

z = -0.49457, p-value = 0.6209

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau
-0.04062164

Resultado: P menor que 0.05

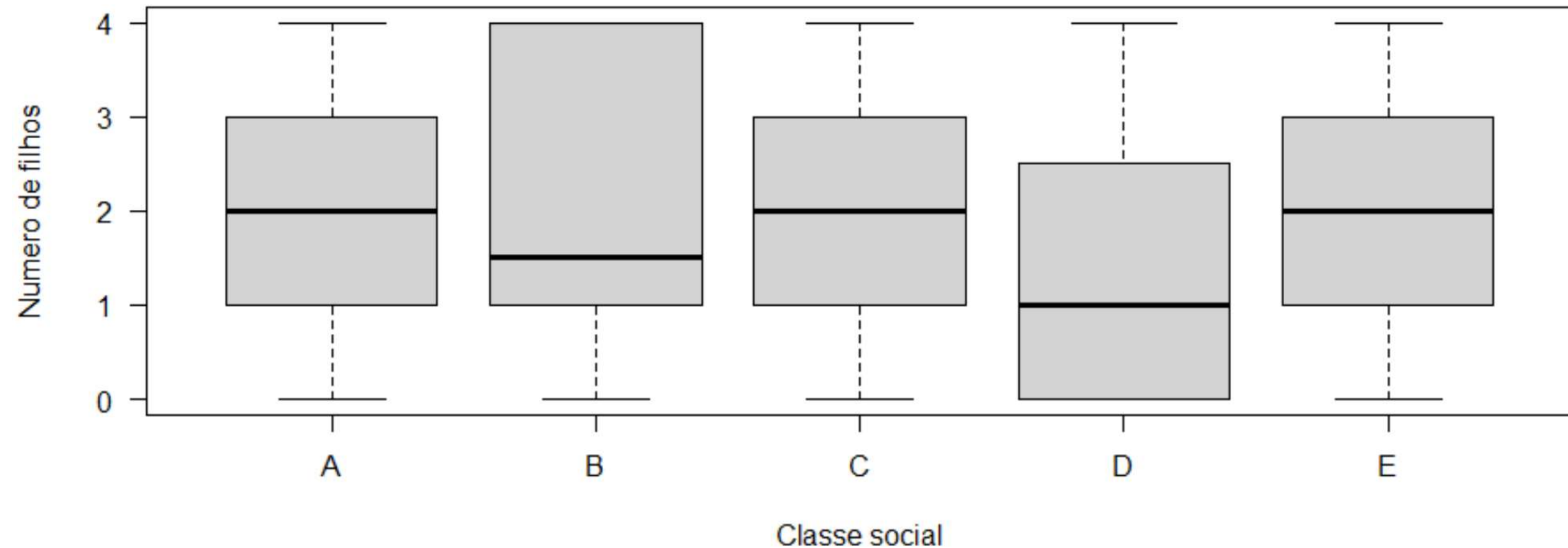
Conclusão: H_0 é verdadeira

Hipóteses testadas

H_0 : Não há correlação entre as variáveis

H_1 : Há (SIM) correlação entre as variáveis

Teste por correlação de Kendall



Regressões lineares

Regressão linear simples

Variável resposta

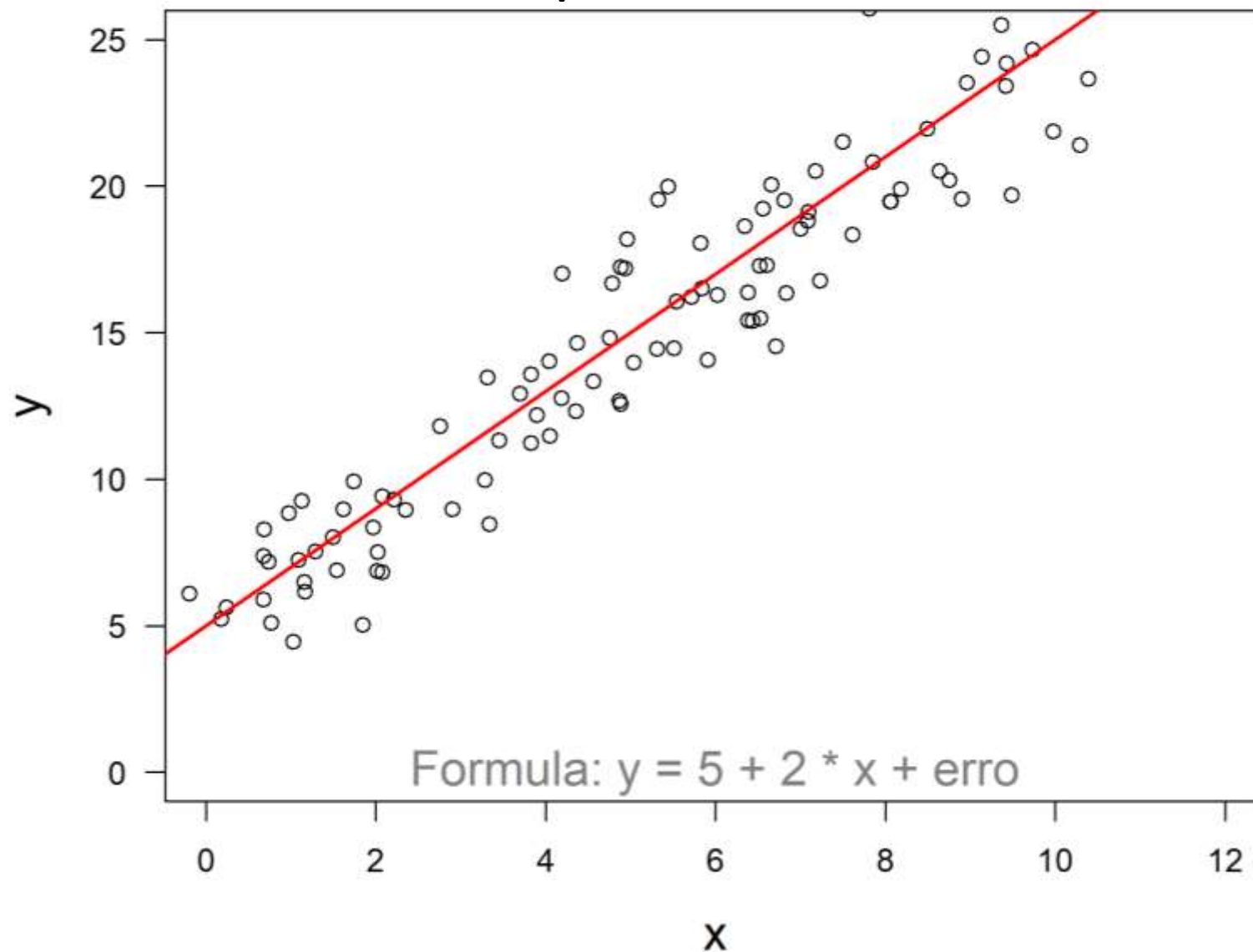
Variável preditora

1 variável
quantitativa

1 variável
quantitativa




Regressão linear simples



Lista 14 pergunta 6

Pergunta 6. Assumindo um nível de significância de 5%, existe o peso de recém-nascidos (Coluna PesoBebe) pode ser previsto pela duração da gestação (Coluna DuracaoGestacao) através de uma modelagem baseada em regressão linear simples?

1  pontos

- ☒ Não existem evidencias de que a variável peso de recém-nascidos pode ser prevista pela duração da gestação. 
- ☐ Existem evidencias (SIM) de que a variável peso de recém-nascidos pode ser prevista pela duração da gestação.
- ☐ Não consegui responder essa perguntar porque não entendi como aplicar o teste.
- ☐ Não consegui responder essa perguntar porque não tenho computador disponível para executar o teste.
- ☐ Não sei/Não quero responder essa pergunta.

Script

```
# Exercício 6
```

```
# peso do bebe ~ duração da gestação
```

```
modelo<-lm(tabela1$PesoBebe~tabela1$DuracaoGestacao)
```

```
summary(modelo)
```

```
# Gráfico de dispersão
```

```
plot(tabela1$PesoBebe~tabela1$DuracaoGestacao, xlab="Duração da gestação (em dias)",ylab="Peso do bebê")
```

Regressão linear simples

Resultado: P menor que 0.05
Conclusão: H_0 é verdadeira

```
> summary(modelo)
```

```
Call:
lm(formula = tabela1$PesoBebe ~ tabela1$DuracaoGestacao)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-505.55	-161.72	6.18	175.32	712.39

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4081.801	1381.989	2.954	0.00393 **
tabela1\$DuracaoGestacao	-3.866	4.938	-0.783	0.43556

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

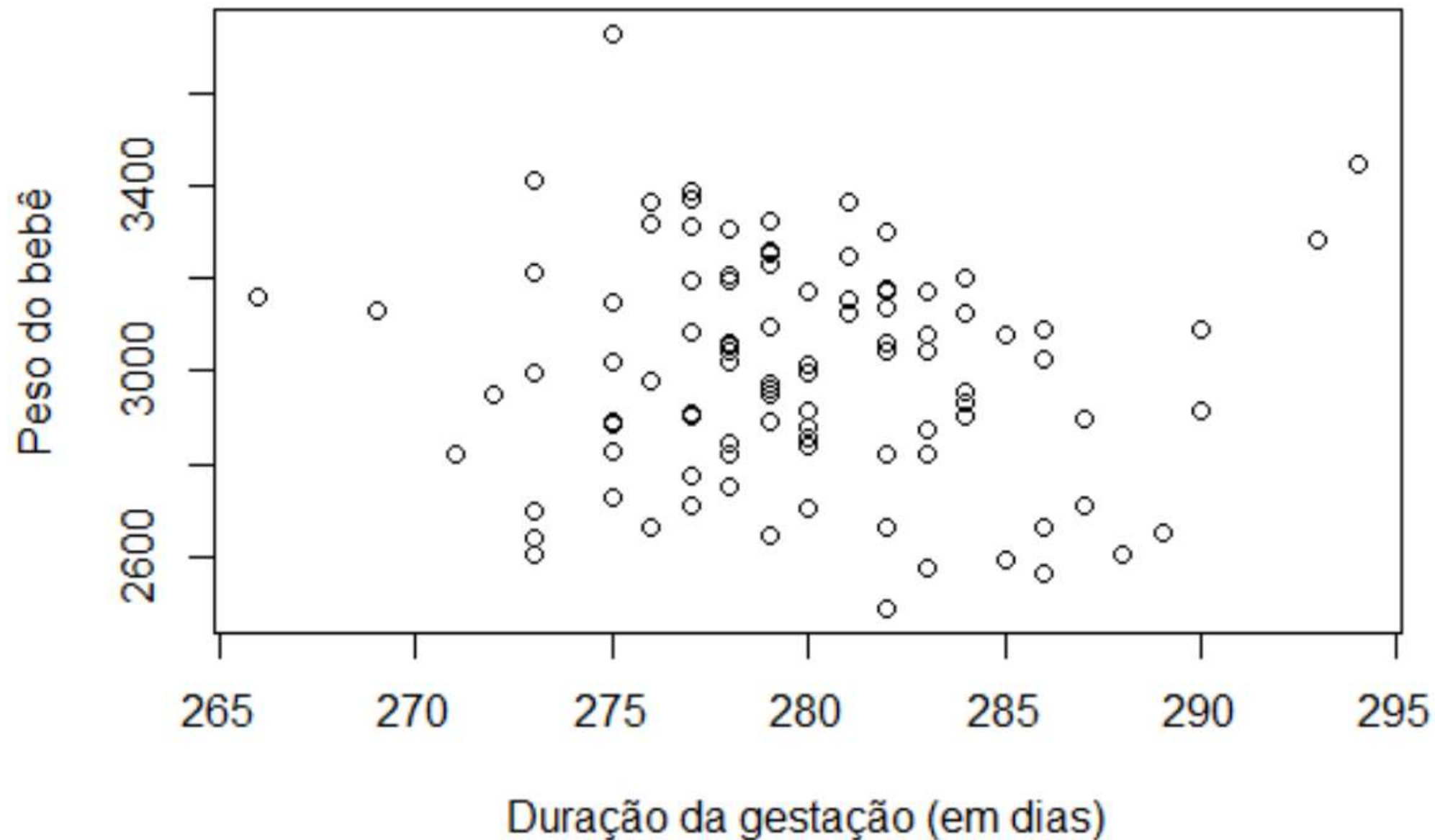
```
Residual standard error: 237.9 on 98 degrees of freedom
Multiple R-squared:  0.006216,    Adjusted R-squared:  -0.003925
F-statistic: 0.613 on 1 and 98 DF,  p-value: 0.4356
```

Hipóteses testadas

H_0 : β_1 não é diferente de zero. Portanto, não há associação entre as variáveis.


H_1 : β_1 é diferente de zero. Portanto, há (sim) associação entre as variáveis.


Gráfico de dispersão



Lista 14 pergunta 7

Pergunta 7. Assumindo um nível de significância de 5%, existe o peso de recém-nascidos (Coluna PesoBebe) pode ser previsto pela altura da mãe (Coluna AlturaMae) através de uma modelagem baseada em regressão linear simples?

1  pontos

- ☐ Não existem evidencias de que a variável peso de recém-nascidos pode ser prevista pela altura da mãe.
- ☒ Existem evidencias (SIM) de que a variável peso de recém-nascidos pode ser prevista pela altura da mãe. 
- ☐ Não consegui responder essa perguntar porque não entendi como aplicar o teste.
- ☐ Não consegui responder essa perguntar porque não tenho computador disponível para executar o teste.
- ☐ Não sei/Não quero responder essa pergunta.

Script

```
# Exercício 7
```

```
modelo<-lm(tabela1$PesoBebe~tabela1$AlturaMae)
```

```
summary(modelo)
```

```
# Gráfico de dispersão
```

```
plot(tabela1$PesoBebe~tabela1$AlturaMae, xlab="Altura da mãe (cm)",ylab="Peso  
do bebê")
```

```
abline(modelo, col="red", lwd=2)
```


Regressão linear simples

Resultado: P menor que 0.05
Conclusão: H_1 é verdadeira

Hipóteses testadas

H_0 : β_1 não é diferente de zero. Portanto, não há associação entre as variáveis.

H_1 : β_1 é diferente de zero. Portanto, há (sim) associação entre as variáveis.

```
> summary(modelo)
```

```
Call:
lm(formula = tabela1$PesoBebe ~ tabela1$AlturaMae)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-167.698	-58.620	3.851	45.523	298.541

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-337.6043	120.9051	-2.792	0.00629	**
tabela1\$AlturaMae	20.9448	0.7571	27.666	< 2e-16	***

```
---
```

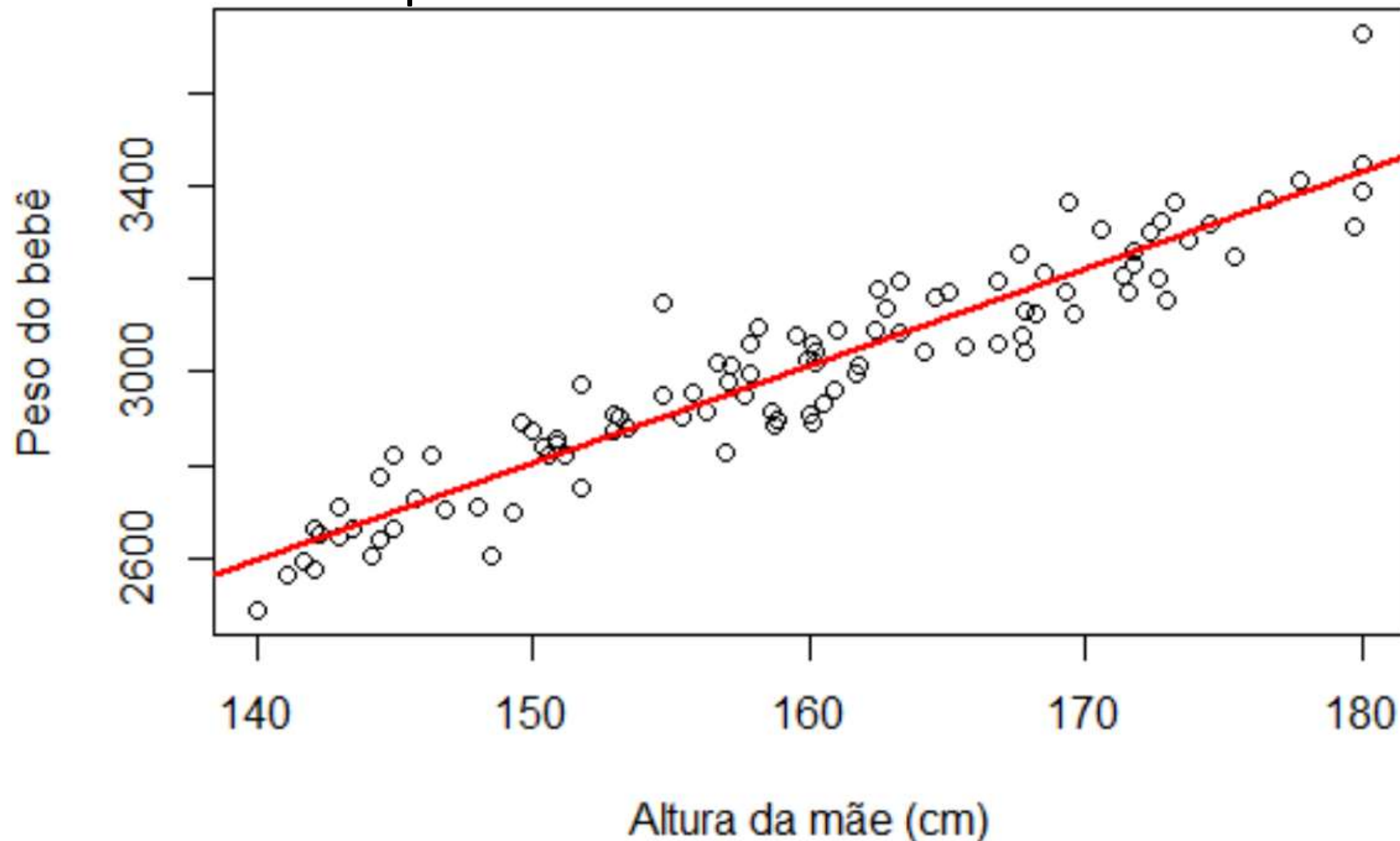
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 80.4 on 98 degrees of freedom
```

```
Multiple R-squared:  0.8865, Adjusted R-squared:  0.8853
```

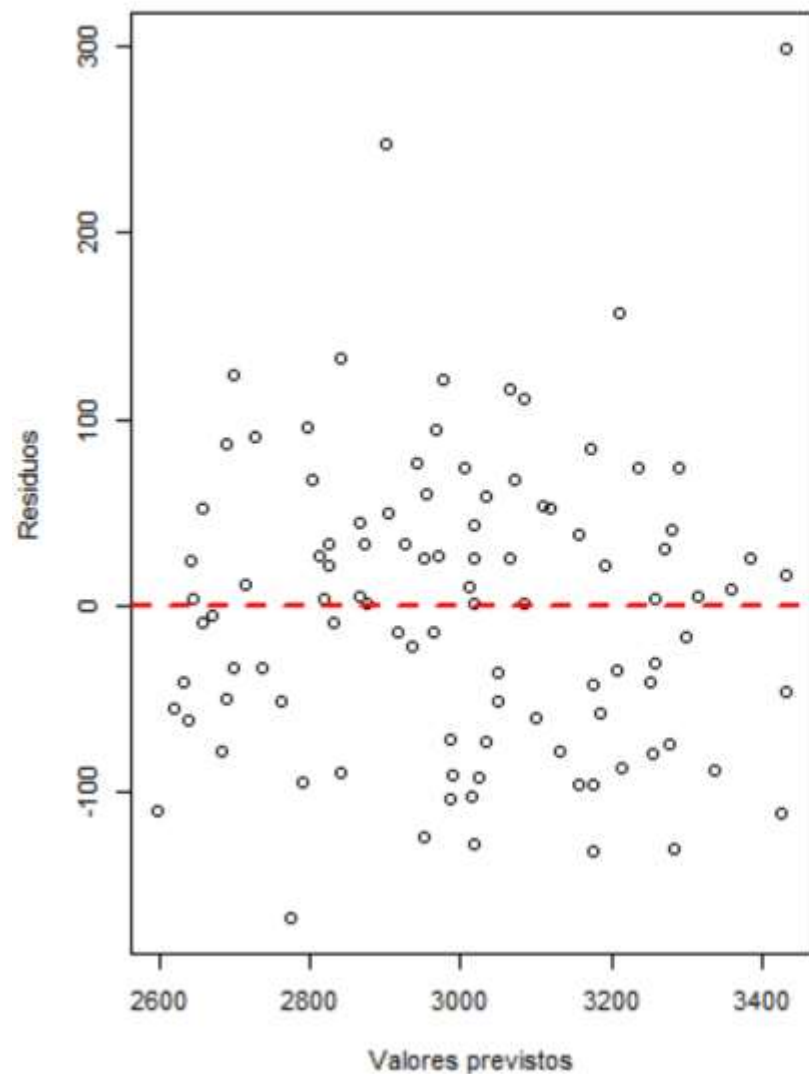
```
F-statistic: 765.4 on 1 and 98 Df, p-value: < 2.2e-16
```

Gráfico de dispersão

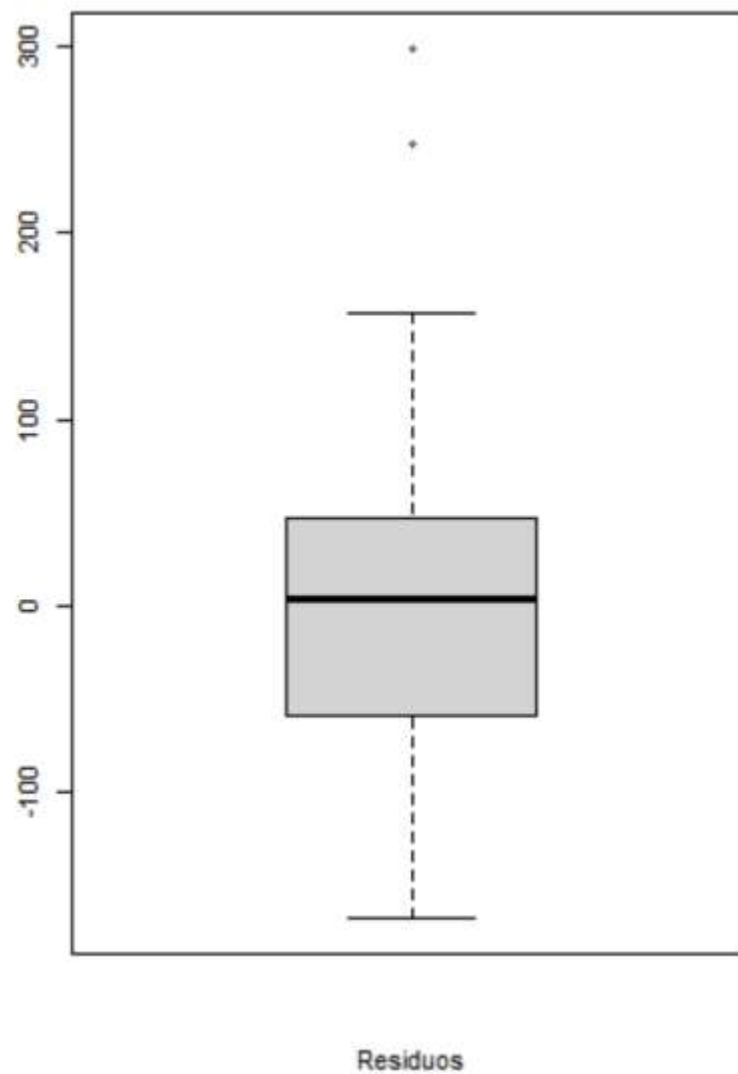


Análise de resíduos

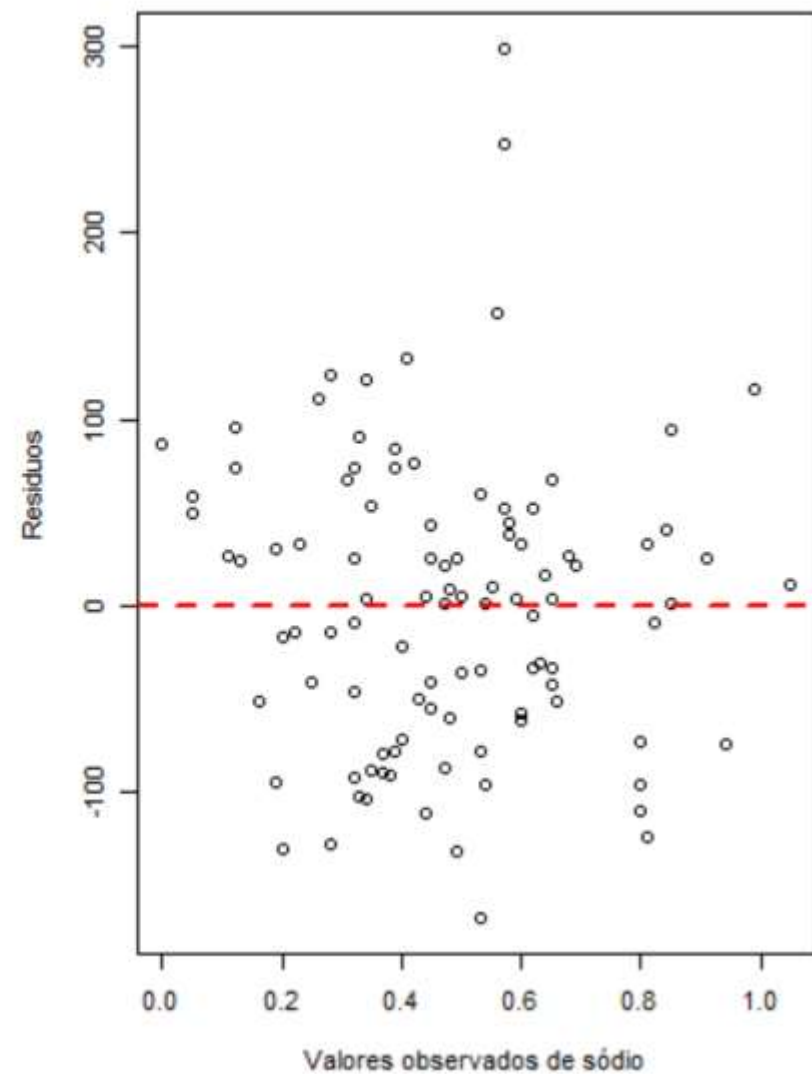
Resíduos vs valores previstos



Boxplot dos resíduos

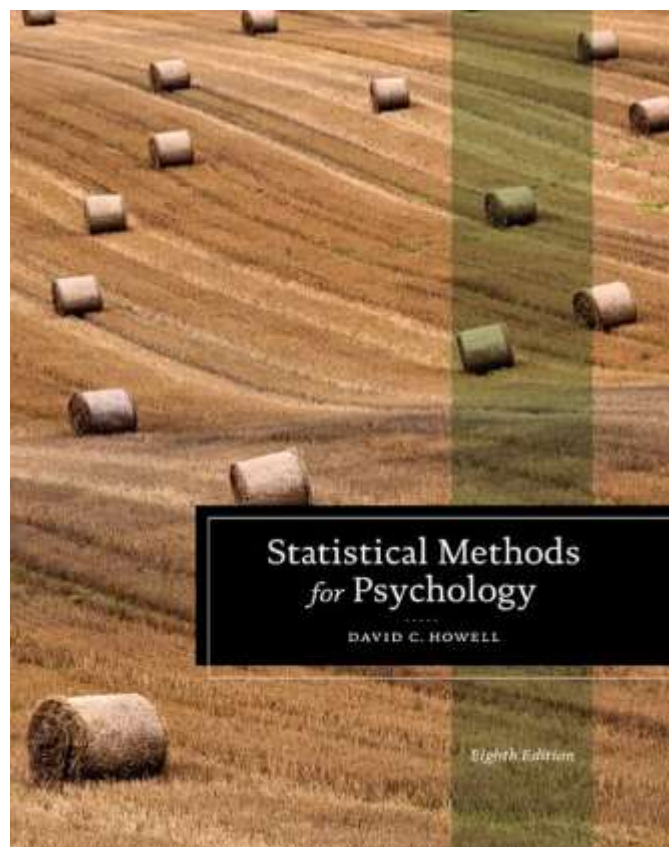


Resíduos vs. variável explanatória



Aula de hoje

Literatura recomendada da aula de hoje

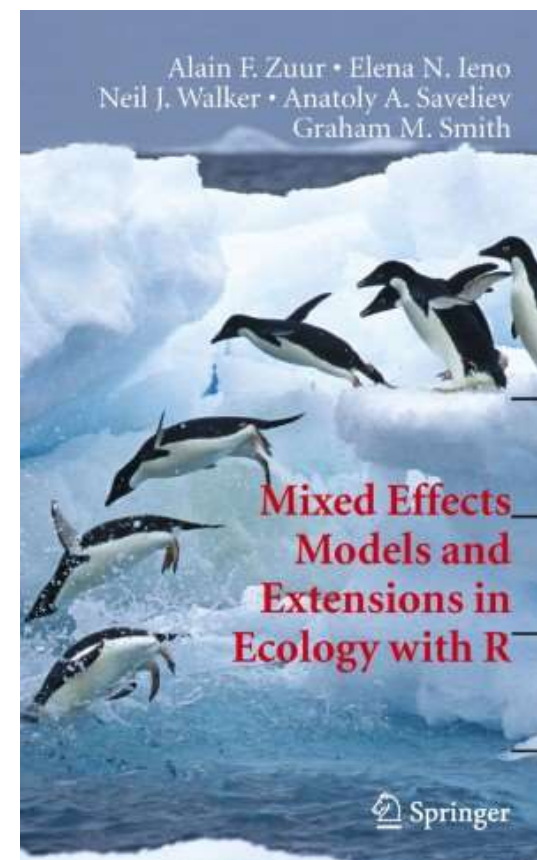


[book] Statistical methods for psychology

DC Howell - 2009 - books.google.com

STATISTICAL METHODS FOR PSYCHOLOGY surveys the statistical techniques commonly used in the behavioral and social sciences, especially psychology and education. To help students gain a better understanding of the specific statistical hypothesis tests that are ...

☆ 𐀀 Cited by 11464 Related articles All 4 versions



[book] Mixed effects models and extensions in ecology with R

A Zuur, EN Ieno, N Walker, [AA Saveliev](#), GM Smith - 2009 - books.google.com

... in Drug Development Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, **Logistic**, Survival, and ... Sciences Zuur/Ieno/Smith: Analysing **Ecological** Data Zuur/Ieno/Walker/Saveliev/Smith: **Mixed Effects** Models and **Extensions** in **Ecology** ...

☆ 𐀀 Cited by 13330 Related articles All 10 versions

Regressões múltiplas

Regressões



- Pense nos modelos matemáticos como manequins, onde você ajusta seus dados a uma estrutura previamente estabelecida

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

Equação que define valores esperados de uma dada variável resposta (y), dado a observação de um conjunto variáveis preditoras ($x_1, x_2, x_3, x_4, \dots$)

Regressão linear múltipla

Variável
resposta

1 variável
quantitativa
ou
qualitativa

Variável
preditora 1

variável
quantitativa
ou
qualitativa

+

Variável
preditora 2

+

variável
quantitativa
ou
qualitativa

+ ...

+ ...



Porque fazer uma regressão múltipla em vez de varias simples?

- Muitas vezes ao realizar varias regressões simples, você pode ser levado ao erro por desconsiderar como a combinação das suas variáveis resposta afeta sua variável preditor

Exemplo para ilustrar a importância de regressões múltiplas

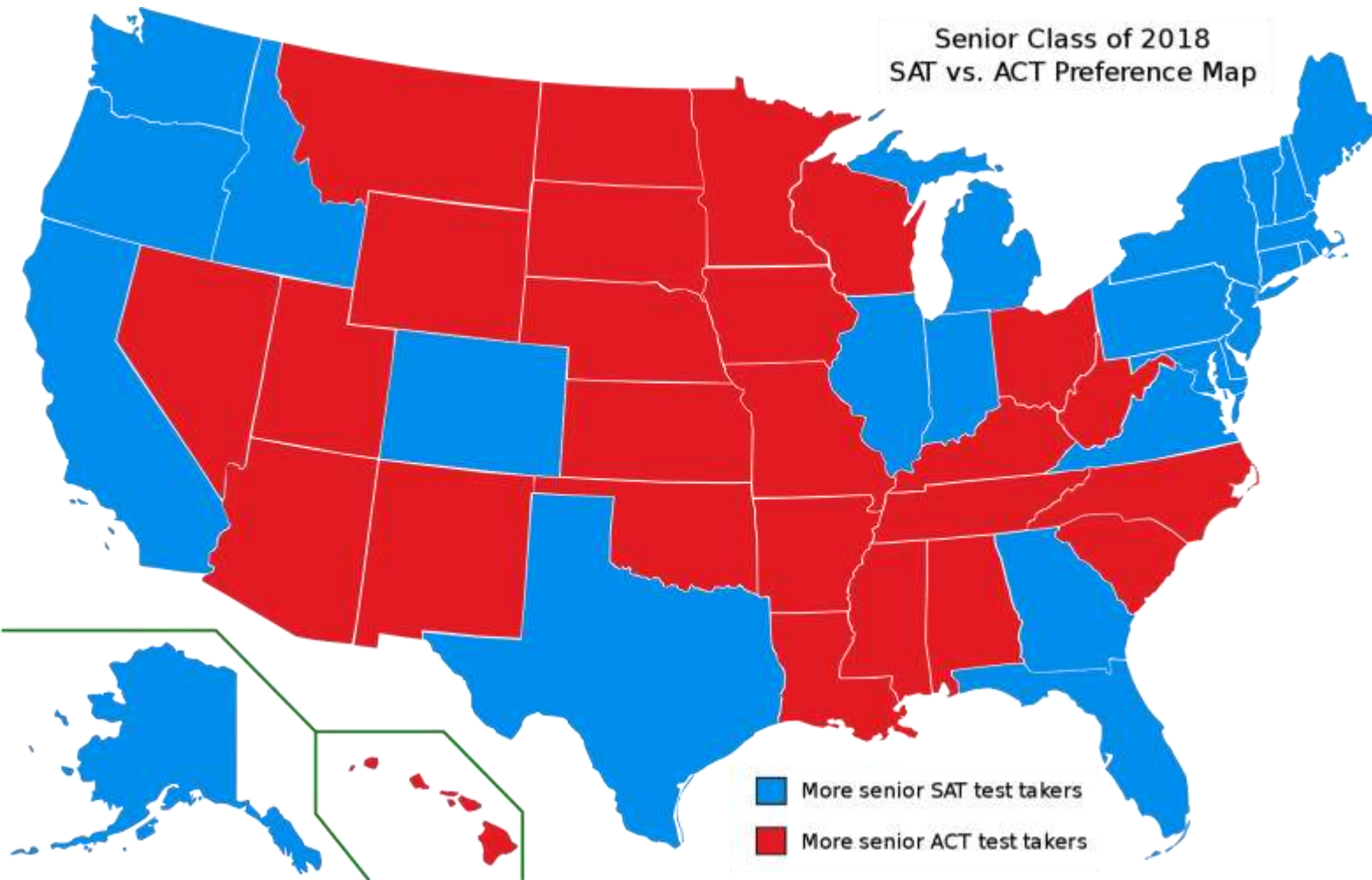
- Pergunta: A quantidade de dinheiro investido em educação está ligado ao desempenho de alunos em um teste de conhecimento?
- Exemplo extraído de: Howell 2012

Exemplo para ilustrar a importância de regressões múltiplas



- Nos EUA existem 2 exames de desempenho de interessados em ingressar em universidades
- ACT
 - Foco em conhecimento teórico retido
 - Usado em um grande número de universidades, por todo EUA
- SAT
 - Foco em habilidades individuais
 - Usado nas universidades mais renomadas dos EUA, frequentemente localizadas na região Norte e Oeste dos EUA

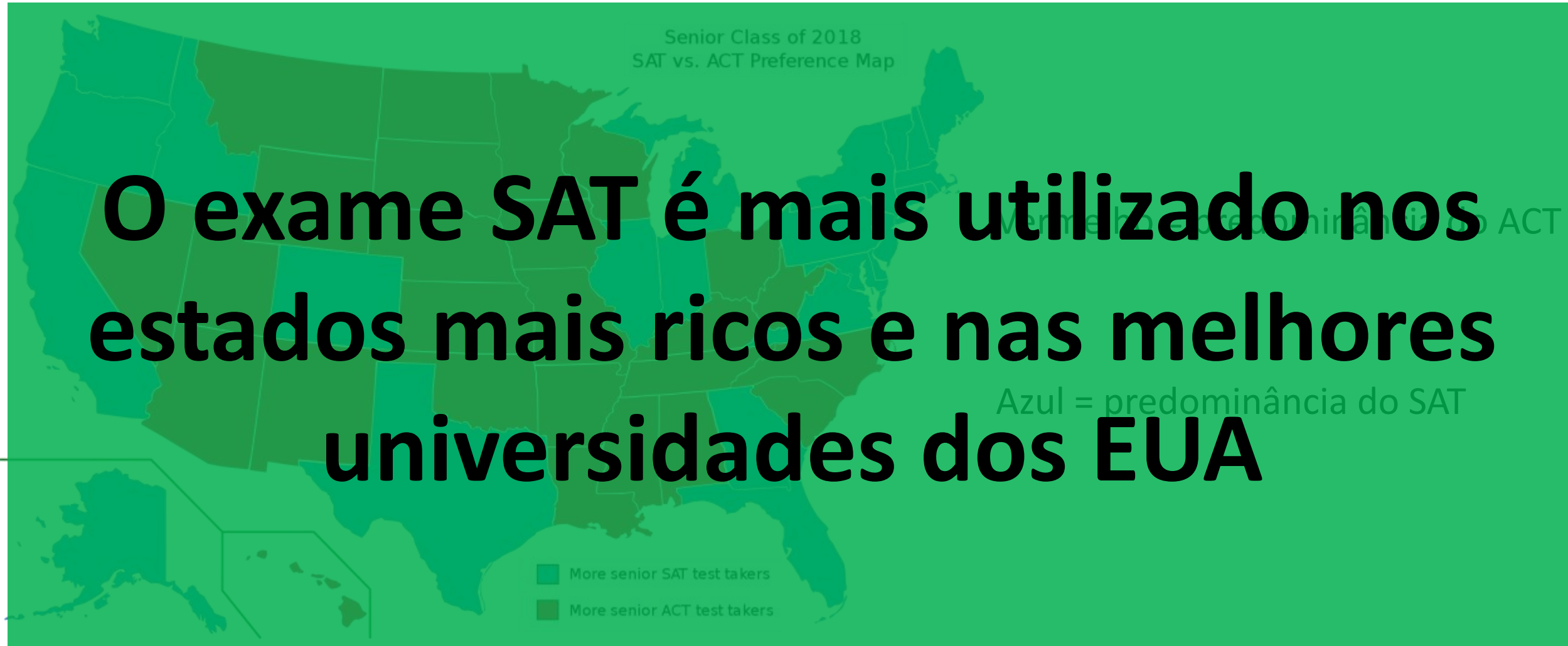
Exemplo para ilustrar a importância de regressões múltiplas



Vermelho = predominância do ACT

Azul = predominância do SAT

Exemplo para ilustrar a importância de regressões múltiplas



Exemplo para ilustrar a importância de regressões múltiplas

Nome da universidade	Prêmios Nobel		SAT			
	Número de prêmios Nobel	Colocação ranking mundial de prêmios Nobel	Leitura e escrita (máx. 800)	Matemática (máx. 800)	Nota exigida (máx. 1600)	% de acerto exigido
Harvard University	161	1	740	770	1510	94,38
University of Chicago	100	4	750	780	1530	95,63
Massachusetts Institute of Technology (MIT)	97	5	741	790	1531	95,69
Columbia University	96	6	740	770	1510	94,38
Stanford University	86	7	740	760	1500	93,75
California Institute of Technology (Caltech)	76	8	760	795	1555	97,19
Princeton University	69	10	740	770	1510	94,38
Yale University	65	11	750	760	1510	94,38
Cornell University	61	12	720	750	1470	91,88
Johns Hopkins University	39	17	735	780	1515	94,69
University of Pennsylvania	36	21	730	770	1500	93,75
Washington University in St. Louis	25	30	735	780	1515	94,69

Ajustando a pergunta

- Pergunta inicial: A quantidade de dinheiro investido em educação está ligado ao desempenho de alunos em um teste de conhecimento?
- Vamos focar no SAT que é o exame relacionado as universidade de maior prestígio dos EUA
- Traduzindo para estatística: o desempenho dos alunos pode ser previsto pela quantidade de dinheiro gasto na educação?

$\text{NotaSAT} \sim \text{GastoEducação}$

Conjunto de dados

Table 15.1 Data on performance versus expenditures on education

State	Expend	PTratio	Salary	PctSAT	SAT	PctACT	ACT
Alabama	4.405	17.2	31.144	8	1029	61	20.2
Alaska	8.963	17.6	47.951	47	934	32	21.0
Arizona	4.778	19.3	32.175	27	944	27	21.1
Arkansas	4.459	7.1	28.934	6	1005	66	20.3
California	4.992	24.0	41.078	45	902	11	21.0
Colorado	5.443	18.4	34.571	29	980	62	21.5
Connecticut	8.817	14.4	50.045	81	908	3	21.4
Delaware	7.030	16.6	39.076	68	897	3	21.0
Florida	5.718	19.1	32.588	48	889	36	20.4
Georgia	5.193	16.3	32.291	65	854	16	20.2
Hawaii	6.078	17.9	38.518	57	889	17	21.6
Idaho	4.210	19.1	29.783	15	979	62	21.4
Illinois	6.136	17.3	39.431	13	1048	69	21.2
Indiana	5.826	17.5	36.785	58	882	19	21.2
Iowa	5.483	15.8	31.511	5	1099	64	22.1
Kansas	5.817	15.1	34.652	9	1060	74	21.7
Kentucky	5.217	17.0	32.257	11	999	65	20.1
Louisiana	4.761	16.8	26.461	9	1021	80	19.4
Maine	6.428	13.8	31.972	68	896	2	21.5
Maryland	7.245	17.0	40.661	64	909	11	20.7
Massachusetts	7.287	14.8	40.795	80	907	6	21.6
Michigan	6.994	20.1	41.895	11	1033	68	21.3
Minnesota	6.000	17.5	35.948	9	1085	60	22.1
Mississippi	4.080	17.5	26.818	4	1036	79	18.7
Missouri	5.383	15.5	31.189	9	1045	64	21.5

Table 15.1 (Continued)

State	Expend	PTratio	Salary	PctSAT	SAT	PctACT	ACT
Montana	5.692	16.3	28.785	21	1009	55	21.9
Nebraska	5.935	14.5	30.922	9	1050	73	21.7
Nevada	5.160	18.7	34.836	30	917	39	21.3
New Hampshire	5.859	15.6	34.720	70	935	4	22.3
New Jersey	9.774	13.8	46.087	70	898	3	20.8
New Mexico	4.586	17.2	28.493	11	1015	59	20.3
New York	9.623	15.2	47.612	74	892	16	21.9
North Carolina	5.077	16.2	30.793	60	865	11	19.3
North Dakota	4.775	15.3	26.327	5	1107	78	21.4
Ohio	6.162	16.6	36.802	23	975	60	21.3
Oklahoma	4.845	15.5	28.172	9	1027	66	20.6
Oregon	6.436	19.9	38.555	51	947	12	22.3
Pennsylvania	7.109	17.1	44.510	70	880	8	21.0
Rhode Island	7.469	14.7	40.729	70	888	2	21.4
South Carolina	4.797	16.4	30.279	58	844	13	18.9
South Dakota	4.775	14.4	25.994	5	1068	68	21.3
Tennessee	4.388	18.6	32.477	12	1040	83	19.7
Texas	5.222	15.7	31.223	47	893	30	20.2
Utah	3.656	24.3	29.082	4	1076	69	21.5
Vermont	6.750	13.8	35.406	68	901	7	21.9
Virginia	5.327	14.6	33.987	65	896	6	20.7
Washington	5.906	20.2	36.151	48	937	16	22.4
West Virginia	6.107	14.8	31.944	17	932	57	20.0
Wisconsin	6.930	15.9	37.746	9	1073	64	22.3
Wyoming	6.160	14.9	31.285	10	1001	70	21.4

Conjunto de dados

Table 15.1 Data on performance versus expenditures on education

State	Expend	PTratio	Salary	PctSAT	SAT	PctACT	ACT
Alabama	4.405	17.2	31.444	8	1029	61	20.2
Alaska	5.3	17.3	41.1	11	1033	68	21.3
Arizona	5.9	15.6	34.720	70	935	4	22.3
Arkansas	4.2	17.2	31.444	8	1029	61	20.2
California	7.287	14.8	40.795	80	907	6	21.6
Colorado	6.994	20.1	41.895	11	1033	68	21.3
Connecticut	6.000	17.5	35.948	9	1085	60	22.1
Delaware	4.080	17.5	26.818	4	1036	79	18.7
Florida	5.383	15.5	31.189	9	1045	64	21.5
Georgia	5.383	15.5	31.189	9	1045	64	21.5
Hawaii	5.383	15.5	31.189	9	1045	64	21.5
Idaho	5.383	15.5	31.189	9	1045	64	21.5
Illinois	5.383	15.5	31.189	9	1045	64	21.5
Indiana	5.383	15.5	31.189	9	1045	64	21.5
Iowa	5.383	15.5	31.189	9	1045	64	21.5
Kansas	5.383	15.5	31.189	9	1045	64	21.5
Kentucky	5.383	15.5	31.189	9	1045	64	21.5
Louisiana	5.383	15.5	31.189	9	1045	64	21.5
Maine	5.383	15.5	31.189	9	1045	64	21.5
Massachusetts	7.287	14.8	40.795	80	907	6	21.6
Michigan	6.994	20.1	41.895	11	1033	68	21.3
Minnesota	6.000	17.5	35.948	9	1085	60	22.1
Mississippi	4.080	17.5	26.818	4	1036	79	18.7
Missouri	5.383	15.5	31.189	9	1045	64	21.5

Table 15.1 (Continued)

State	Expend	PTratio	Salary	PctSAT	SAT	PctACT	ACT
Montana	5.692	16.3	28.785	21	1009	55	21.9
Nevada	5.383	15.5	31.189	9	1045	64	21.5
New Hampshire	5.859	15.6	34.720	70	935	4	22.3
New Jersey	9.774	13.8	35.087	70	898	3	20.8
New Mexico	5.383	15.5	31.189	9	1045	64	21.5
New York	9.623	15.2	47.612	74	892	16	21.9
North Carolina	5.077	16.2	30.793	60	865	11	19.3
Ohio	6.162	16.6	36.802	23	975	60	21.3
Oklahoma	4.845	15.5	28.172	9	1027	66	20.6
Pennsylvania	7.109	17.1	44.510	70	880	8	21.0
Rhode Island	7.469	14.7	40.729	70	888	2	21.4
South Dakota	4.775	14.4	25.994	5	1068	68	21.3
Tennessee	4.388	18.6	32.477	12	1040	83	19.7
Texas	5.383	15.5	31.189	9	1045	64	21.5
Utah	5.383	15.5	31.189	9	1045	64	21.5
Vermont	6.750	13.8	35.406	68	901	7	21.9
Virginia	5.383	15.5	31.189	9	1045	64	21.5
Washington	5.383	15.5	31.189	9	1045	64	21.5
West Virginia	6.107	14.8	31.944	17	932	57	20.0
Wisconsin	6.994	20.1	41.895	11	1033	68	21.3
Wyoming	5.383	15.5	31.189	9	1045	64	21.5

- State: Nome do estado americano
- Expend: Gastos com a educação
- PTratio: Razão de alunos / professor
- Salary: Salário dos professores
- PctSAT: Proporção de alunos que fizeram o exame SAT
- SAT: Nota media obtida no exame SAT
- PctACT: Proporção de alunos que fizeram o exame ACT
- ACT: Nota media obtida no exame ACT

Correlação entre as variáveis preditoras

Table 15.2 Correlations between selected variables

		Correlations					
		Expend	PTratio	Salary	PctSAT	SAT	LogPctSAT
Expend	Pearson Correlation	1	−.371**	.870**	.593**	−.381**	.561**
	Sig. (2-tailed)		.008	.000	.000	.006	.000
	N	50	50	50	50	50	50
PTratio	Pearson Correlation	−.371**	1	−.001	−.213	.081	−.132
	Sig. (2-tailed)	.008		.994	.137	.575	.361
	N	50	50	50	50	50	50
Salary	Pearson Correlation	.870**	−.001	1	.617**	−.440**	.613**
	Sig. (2-tailed)	.000	.994		.000	.001	.000
	N	50	50	50	50	50	50
PctSAT	Pearson Correlation	.593**	−.213	.617**	1	−.887**	.961**
	Sig. (2-tailed)	.000	.137	.000		.000	.000
	N	50	50	50	50	50	50
SAT	Pearson Correlation	−.381**	.081	−.440**	−.887**	1	−.926**
	Sig. (2-tailed)	.006	.575	.001	.000		.000
	N	50	50	50	50	50	50
LogPctSAT	Pearson Correlation	.561**	−.132	.613**	.961**	−.926**	1
	Sig. (2-tailed)	.000	.361	.000	.000	.000	
	N	50	50	50	50	50	50

Correlação entre as variáveis preditoras

Table 15.2 Correlations between selected variables

			Correlations				
		Expend	PTratio	Salary	PctSAT	SAT	LogPctSAT
Expend	Pearson Correlation	1	-.371**	.870**	.593**	-.381**	.561**
	Sig. (2-tailed)		.008	.000	.000	.006	.000
	N	50	50	50	50	50	50

Gastos com educação são positivamente correlacionados com:

- Salário dos professores (coluna Salary)
- Proporção de alunos que realizam o exame SAT (colunas PctSAT e Log PctSAT)

Gastos com educação são negativamente correlacionados com:

- Razão de número alunos / número de professores (coluna PTratio)
- Nota obtida no SAT (coluna SAT)

Correlação entre as variáveis preditoras

Table 15.2 Correlations between selected variables

		Correlations					
		Expend	PTratio	Salary	PctSAT	SAT	LogPctSAT
Expend	Pearson Correlation	1	−.371**	.870**	.593**	−.381**	.561**
	Sig. (2-tailed)		.008	.000	.000	.006	.000
	N	50	50	50	50	50	50

Surpreendentemente a correlação de Gastos com Nota, sugere que quanto maior o gasto com educação, pior é a nota dos alunos no exame SAT

Correlação entre as variáveis preditoras

Table 15.2 Correlations between selected variables

		Correlations					
		Expend	PTratio	Salary	PctSAT	SAT	LogPctSAT
Expend	Pearson Correlation	1	-.371**	.870**	.593**	-.381**	.561**
	Sig. (2-tailed)		.008	.000	.000	.006	.000
	N	50	50	50	50	50	50

Se você fizer uma regressão linear essa relação se confirma

Call:

```
lm(formula = SAT ~ Expend)
```

Residuals:

Min	1Q	Median	3Q	Max
-145.074	-46.821	4.087	40.034	128.489

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1089.394	44.399	24.539	<.001
Expend	-20.892	7.328	-2.851	0.00641 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 69.91 on 48 degrees of freedom
Multiple R-squared: 0.1448, Adjusted R-squared: 0.127
F-statistic: 8.128 on 1 and 48 Df, p-value: 0.006408

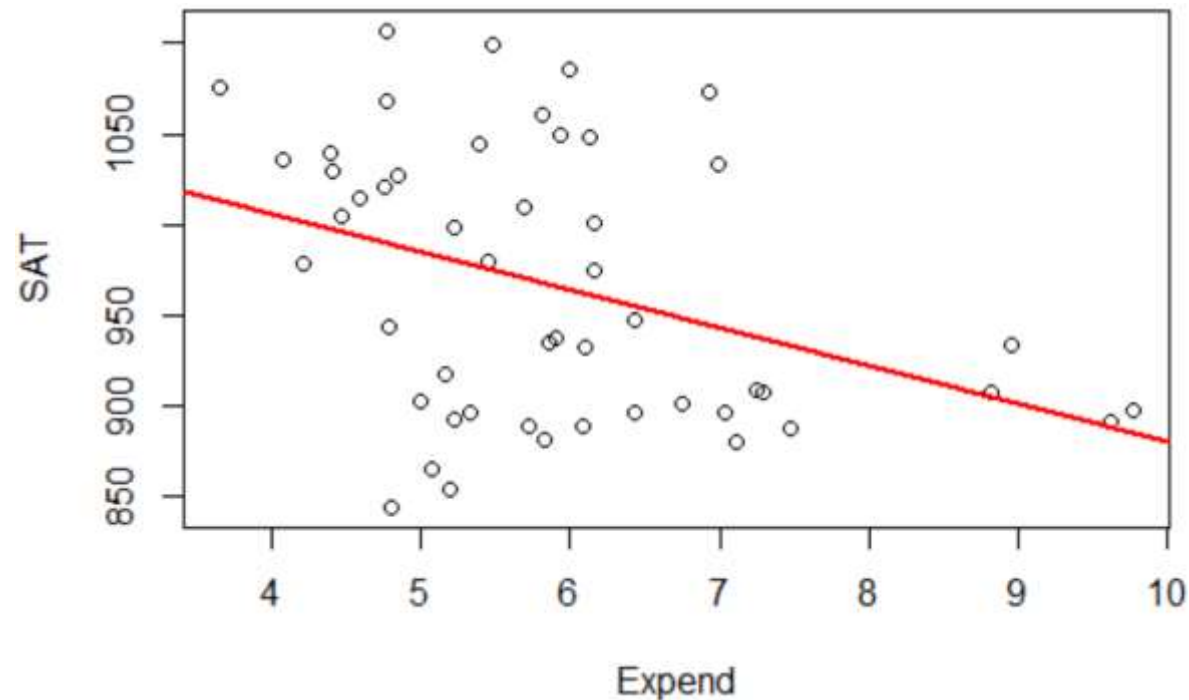
Perceba que β_1
é negativo

Correlação entre as variáveis preditoras

Table 15.2 Correlations between selected variables

		Correlations					
		Expend	PTratio	Salary	PctSAT	SAT	LogPctSAT
Expend	Pearson Correlation	1	-.371**	.870**	.593**	-.381**	.561**
	Sig. (2-tailed)		.008	.000	.000	.006	.000
	N	50	50	50	50	50	50

Se você fizer uma regressão linear essa relação se confirma



Conclusão (momentânea)

- Se você parar por aqui, a conclusão que você teria é:
- Quanto maior o gasto com educação, menor é a nota obtida pelos alunos no exame SAT

Qual é a implicação de assumir essa conclusão?

- Justificativa para cortar dinheiro da educação?
- Desvalorização da escola?
- Diminuir a verba para a educação?

e se você considerar outras
variáveis no seu modelo?

Considerações

- Você viu que nem todos os estados utilizam a avaliação SAT
- SAT é usados nas universidades de maior prestígio
- E se parte da variação da nota de SAT for explicada pela proporção de alunos que realizam o teste?
- Se todos os alunos são obrigados (ou incentivados) a fazer o teste, espera-se uma grande variação da qualidade dos alunos que realizam a prova. Em outras palavras:
 - Muitos alunos fazendo prova = bons e maus alunos fazem a prova
 - Poucos alunos fazendo prova = somente os bons alunos fazem a prova

Traduzindo para realidade brasileira

- No Brasil, existe o estereótipo de que colégios particulares provêm melhor qualidade de ensino que colégios públicos (reforçado pela imprensa).
- Dado que isso influencia na percepção dos alunos quanto a qualidade do seu próprio ensino, podemos esperar que:
 - Alunos de colégio particulares podem se sentir mais seguros para realizar o ENEM
 - Alunos de colégios públicos podem desistir de prestar o ENEM por acharem que vão ir mal na prova

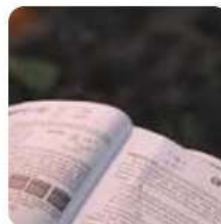


EM Estado de Minas

Próximo Enem deve ser o mais desigual de todos os tempos; entenda

Este ano (no Enem), haverá dois alunos da escola pública", afirma. ... da rede pública não estão tendo o mesmo acesso que os da particular.

1 month ago



FDR - Terra

Enem 2020 marca desigualdade mais forte entre os candidatos; entenda o motivo

Este ano (no Enem), haverá dois alunos da escola pública. Os estudantes da rede pública não estão tendo o mesmo acesso que os da particular ...

1 month ago



Traduzindo para realidade brasileira

- Se de fato isso ocorre, podemos esperar que:
 - Haverá uma grande variabilidade de alunos de colégio particulares realizando o ENEM
 - Apenas os melhores alunos (ou os mais confiantes) de colégios públicos vão realizar o ENEM



Ajustando a pergunta

- Pergunta atualizada: A quantidade de dinheiro gasto na educação em associação com a proporção de alunos que realizam o teste, pode prever a nota obtida no exame SAT?

$\text{NotaSAT} \sim \text{GastoEducação} + \text{ProporçãoTeste}$

Agora você tem em mãos uma
regressão múltipla!

Avaliar pressupostos para regressão

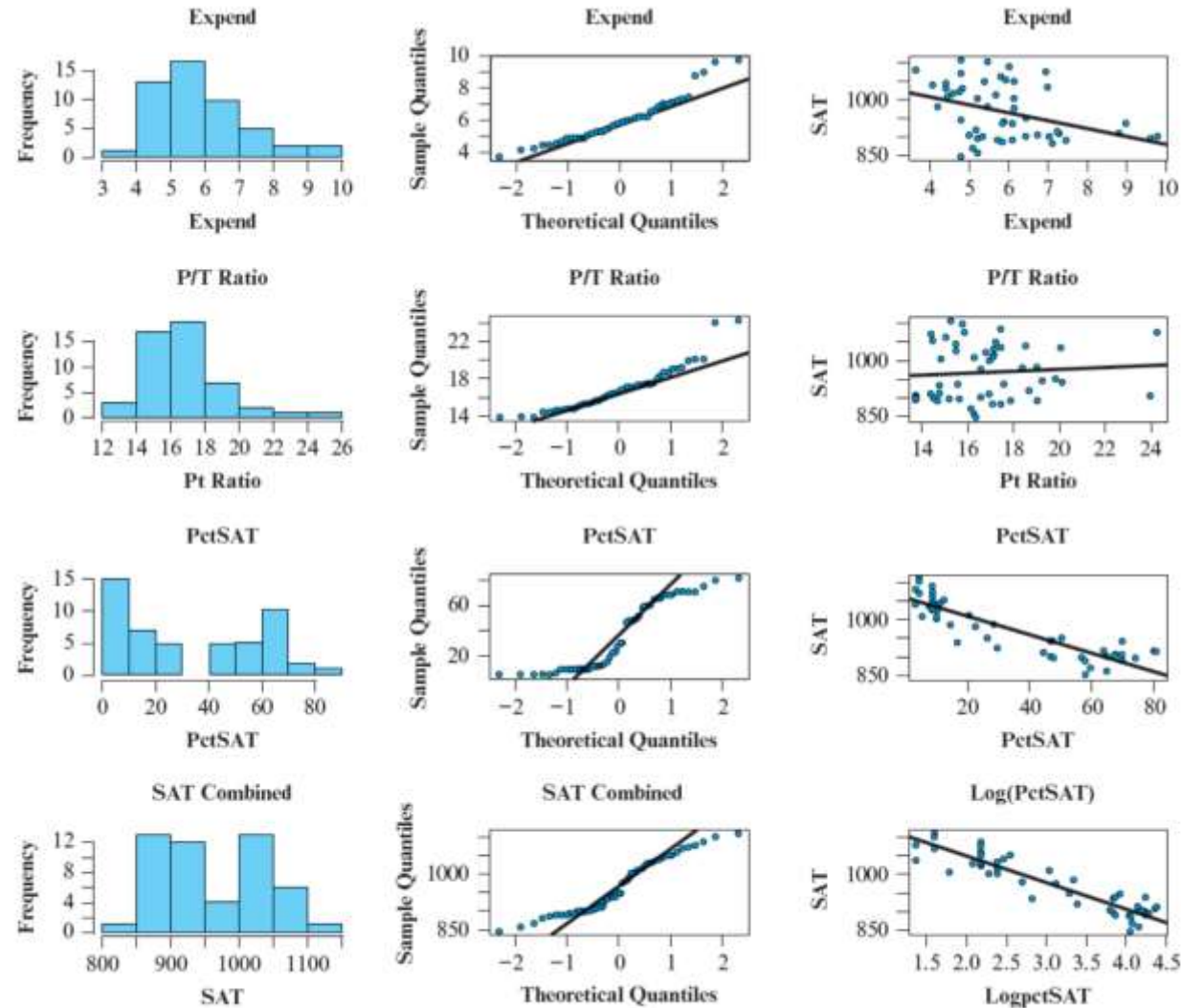


Figure 15.1 Histograms, Q-Q plots, and scatterplots of the variables used in this example

O que esperamos?

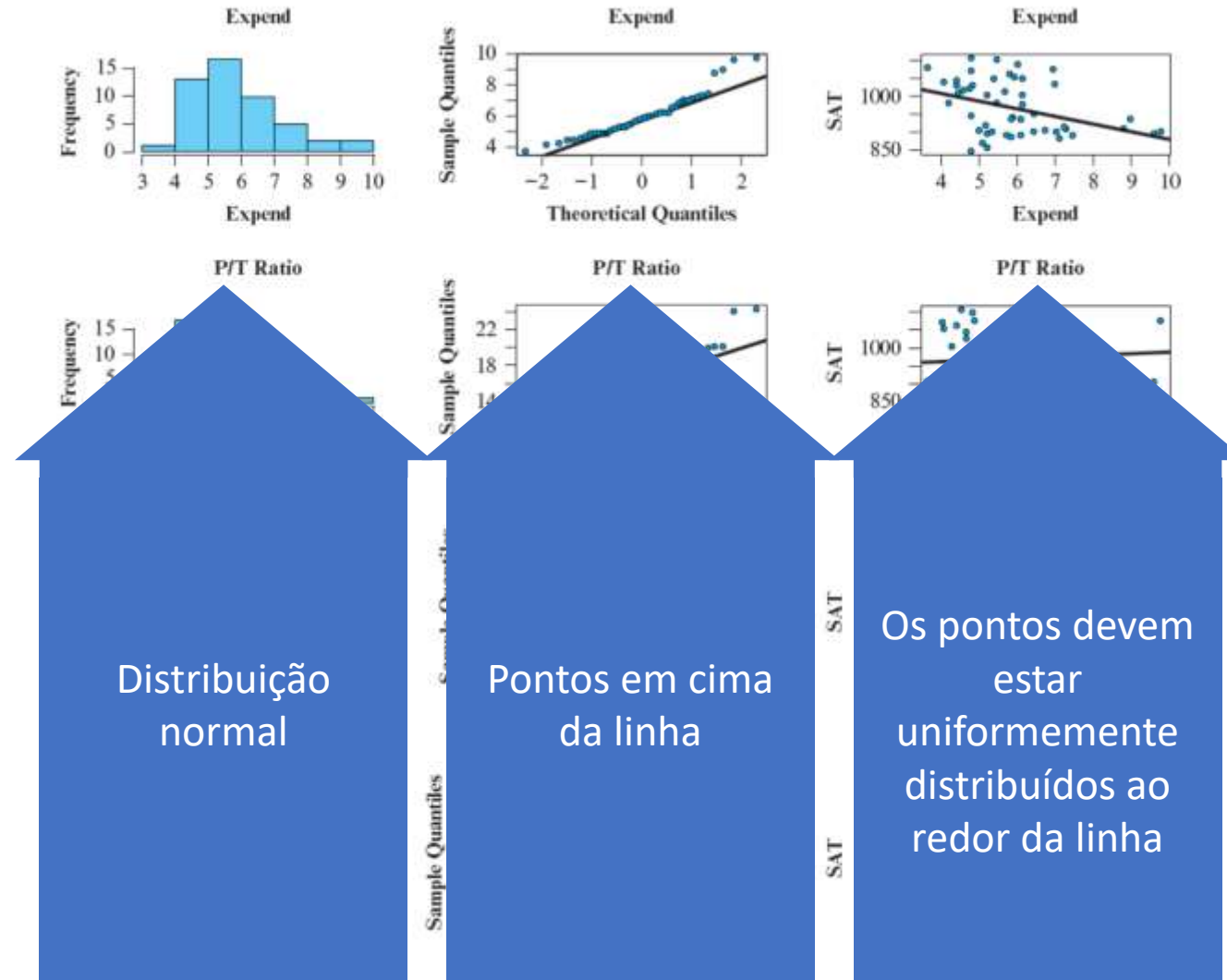


Figure 15.1 Histograms, Q-Q plots, and scatterplots of the variables used in this example

Problemas detectados

Distribuição bimodal

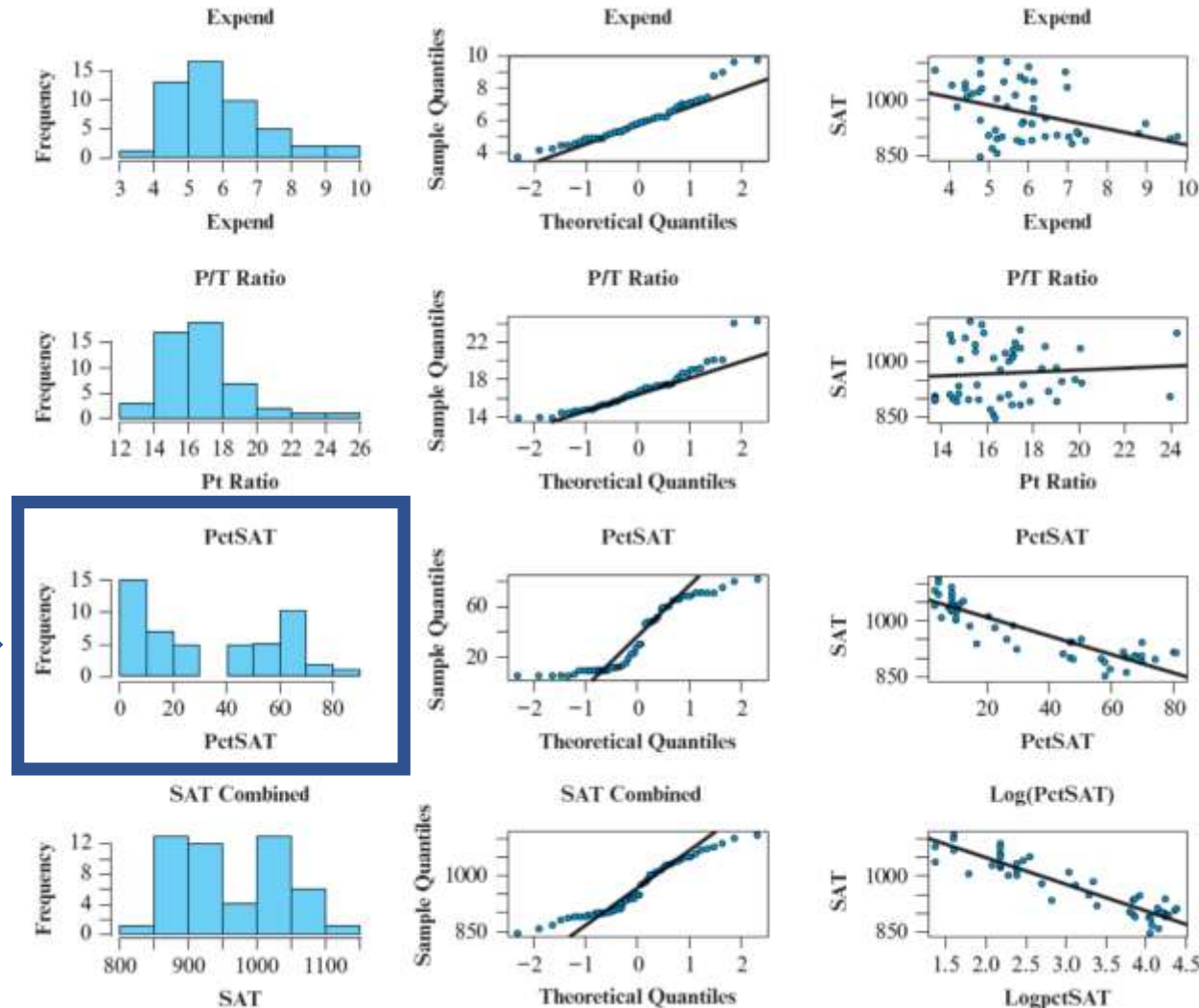
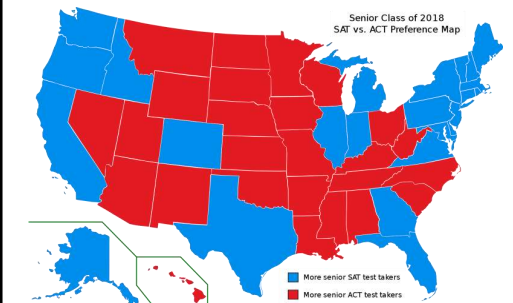


Figure 15.1 Histograms, Q-Q plots, and scatterplots of the variables used in this example

A distribuição bimodal surge porque a proporção de alunos que faz o teste é uma para os estados em azul e outra para os estados em vermelho



Problemas detectados

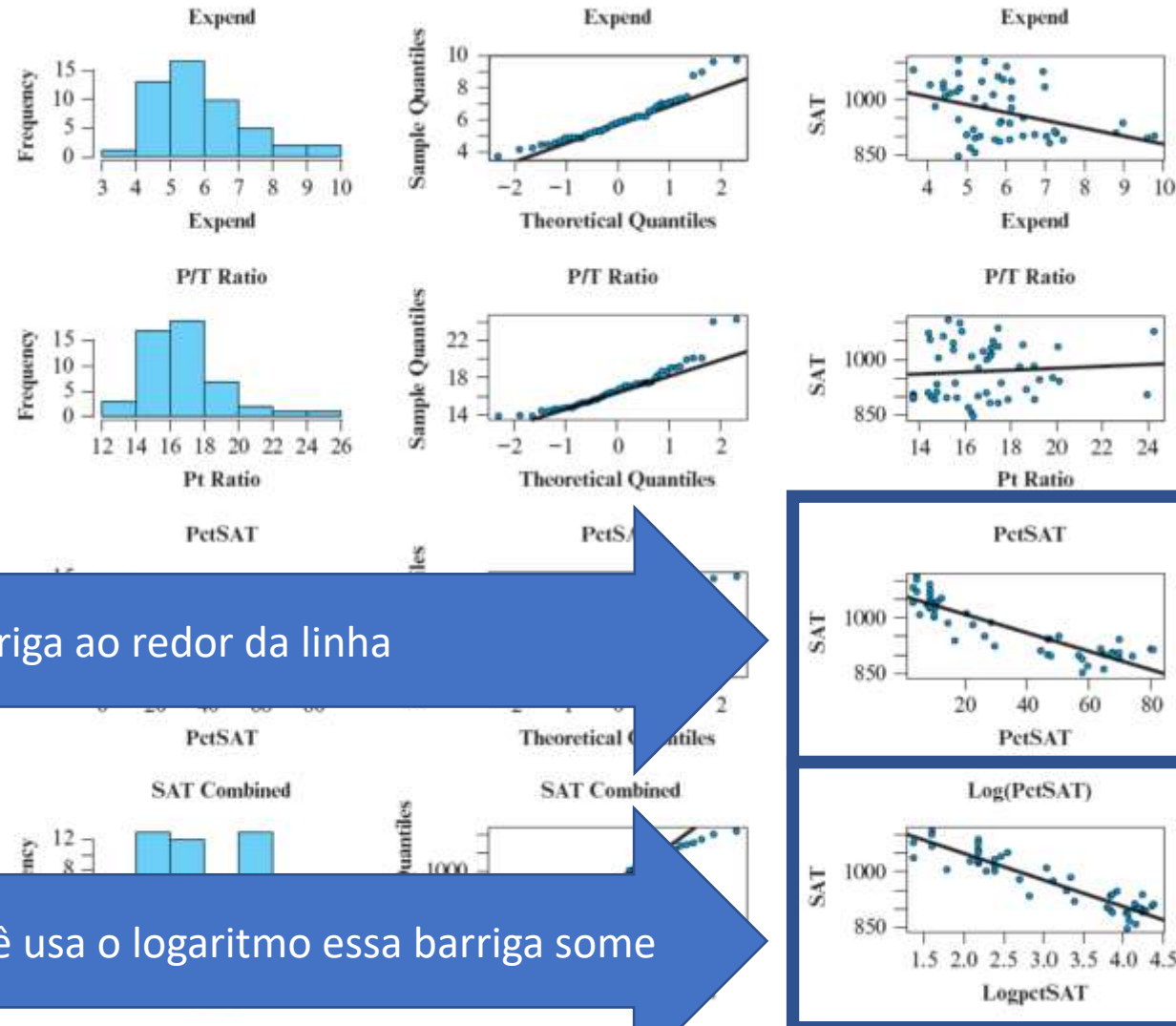


Figure 15.1 Histograms, Q-Q plots, and scatterplots of the variables used in this example

Por esse motivo, vamos trabalhar com o $\text{Log}(\text{PetSAT})$, e não com o dado na escala natural

Aplicando a regressão múltipla

call:

```
lm(formula = SAT ~ Expend + LogPctSAT)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-61.515	-13.616	-2.572	16.541	54.901

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1147.100	16.701	68.684	< 2e-16 ***
Expend	11.129	3.264	3.409	0.00135 **
LogPctSAT	-78.203	4.471	-17.490	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.78 on 47 degrees of freedom

Multiple R-squared: 0.8861, Adjusted R-squared: 0.8813

F-statistic: 182.8 on 2 and 47 DF, p-value: < 2.2e-16

β_1



Aplicando a regressão múltipla

call:

```
lm(formula = SAT ~ Expend + LogPctSAT)
```

Residuals:

Min	1Q	Median	3Q	Max
-61.515	-13.616	-2.572	16.541	54.901

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1147.100	16.701	68.684	< 2e-16	***
Expend	11.129	3.264	3.409	0.00135	**
LogPctSAT	-78.203	4.471	-17.490	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.78 on 47 degrees of freedom

Multiple R-squared: 0.8861, Adjusted R-squared: 0.8813

F-statistic: 182.8 on 2 and 47 DF, p-value: < 2.2e-16

β_2



Aplicando a regressão múltipla

call:

```
lm(formula = SAT ~ Expend + LogPctSAT)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-61.515	-13.616	2.572	16.541	54.901

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1147.100	16.701	68.684	< 2e-16 ***
Expend	11.129	3.264	3.409	0.00135 **
LogPctSAT	-78.203	4.471	-17.490	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.78 on 47 degrees of freedom

Multiple R-squared: 0.8861, Adjusted R-squared: 0.8813

F-statistic: 182.8 on 2 and 47 DF, p-value: < 2.2e-16

Perceba que

- β_1 é positivo
- β_2 é negativo

Note que o R^2
subiu de 0.127
para 0.8813

Relação tridimensional

Gráfico a direita. Cores representam o domínio de cada teste nos estados

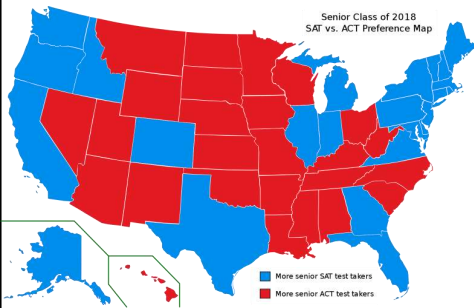
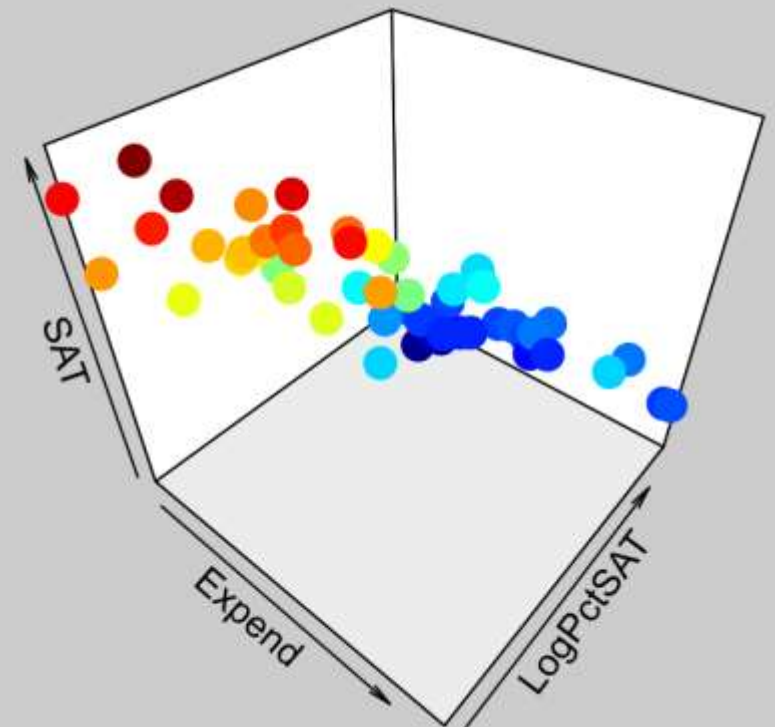
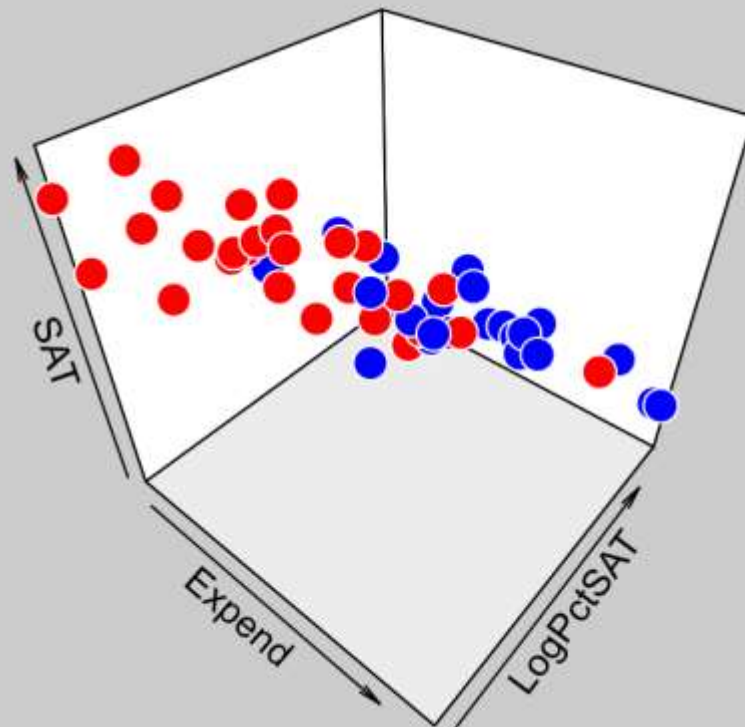
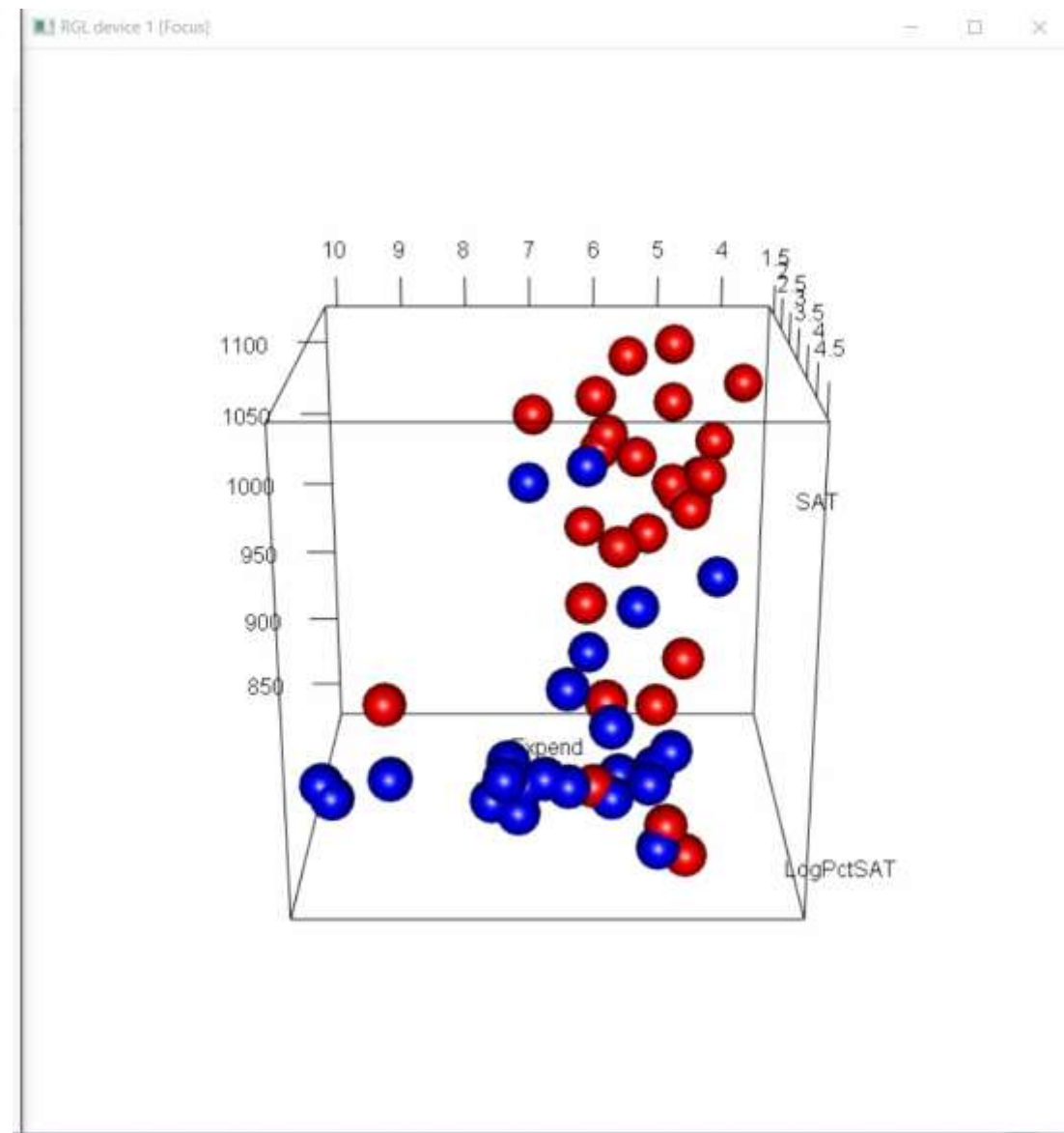


Gráfico a esquerda. Cores representam a intensidade de SAT observada em cada estado

$$\text{SAT} \sim \text{Expend} + \text{LogPctSAT}$$

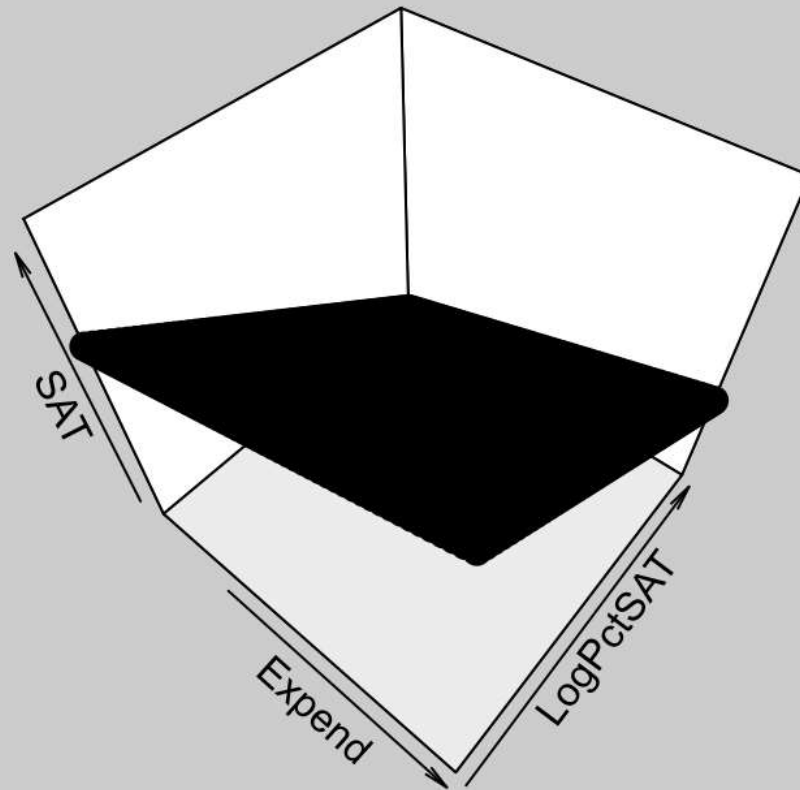


Relação tridimensional

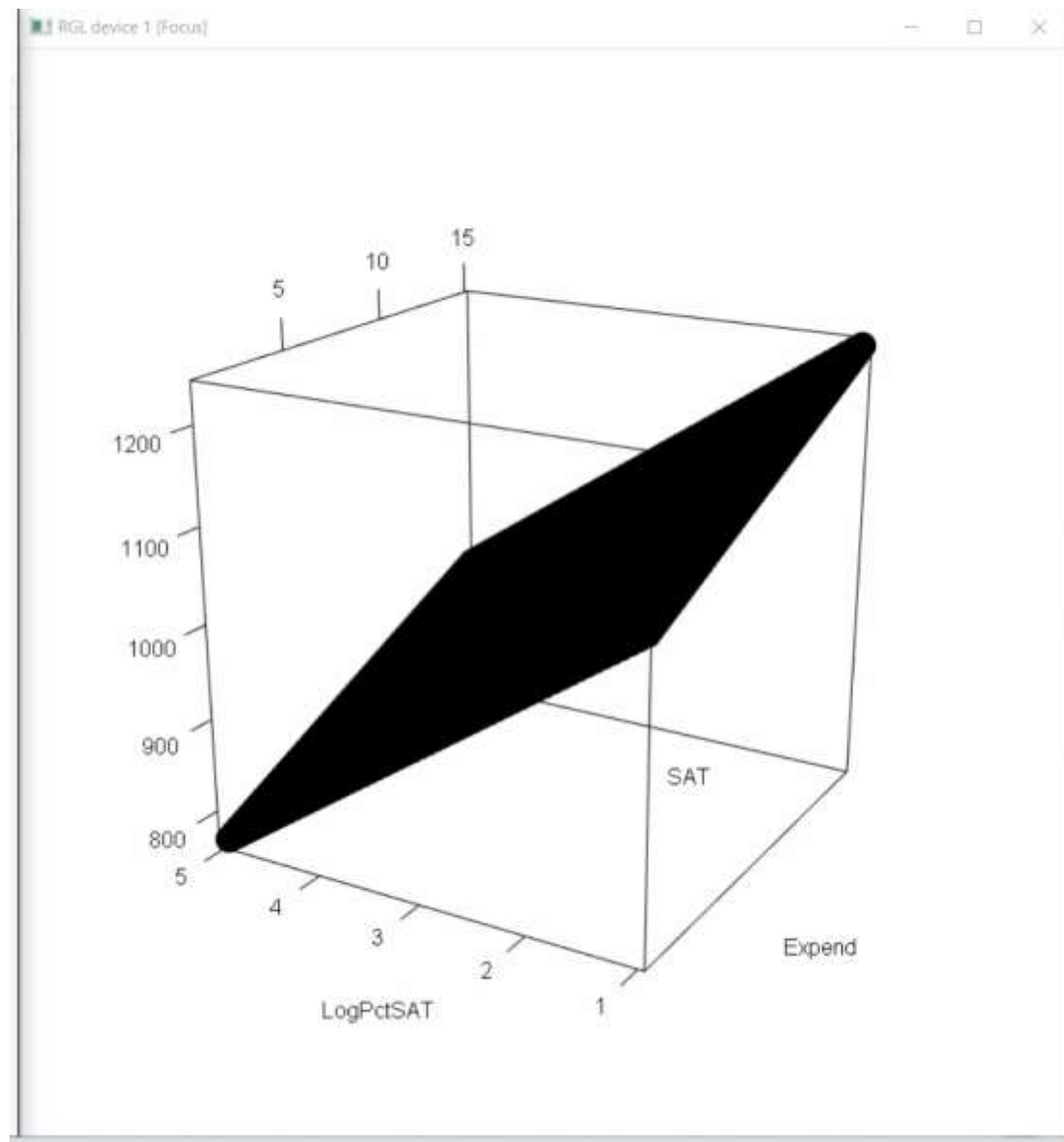


Valores previstos

$$\text{SAT} \sim \text{Expend} + \text{LogPctSAT}$$

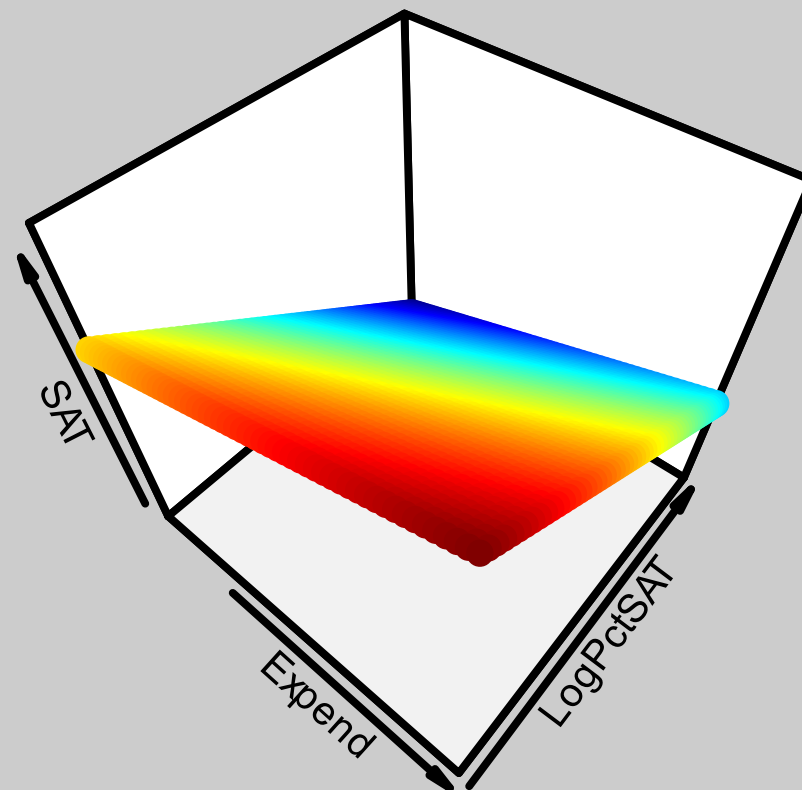


Valores previstos



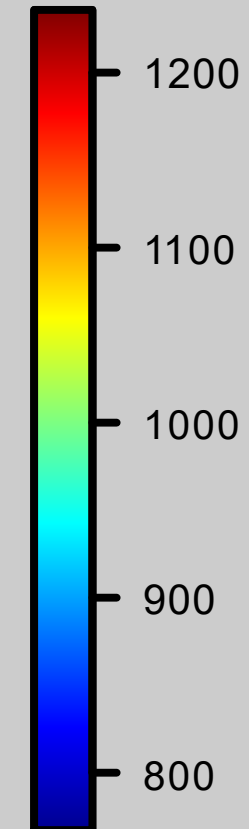
Valores previstos

$$\text{SAT} \sim \text{Expend} + \text{LogPctSAT}$$

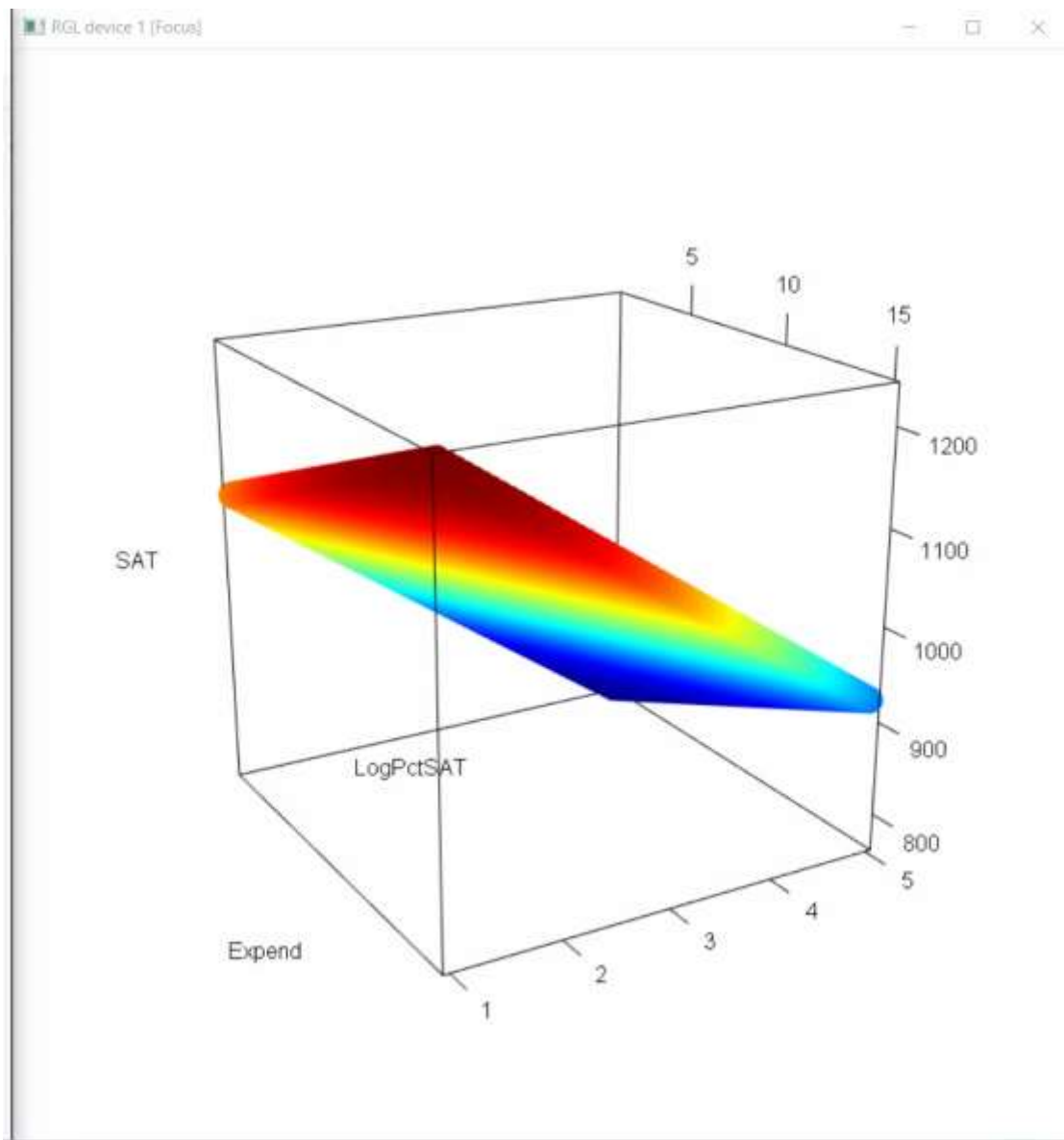


Valores previstos

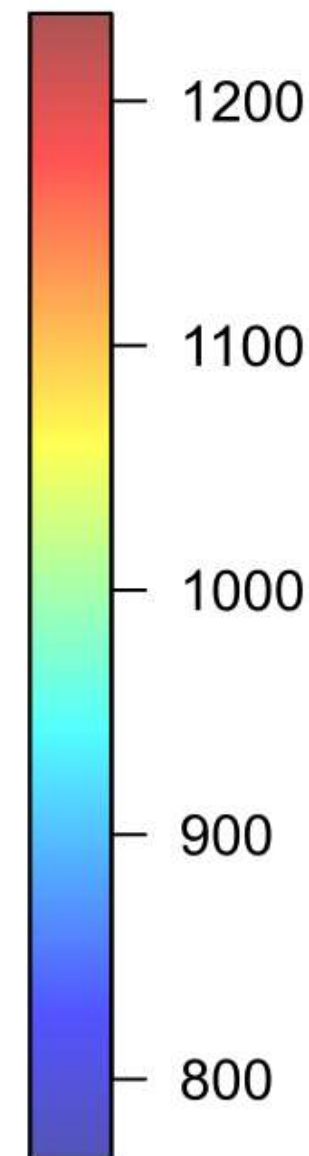
de SAT



Valores previstos



Valores previstos
de SAT



Nota

- Regressões múltiplas são (geralmente) representadas apenas com a tabela de coeficientes
- Gráficos se tornam uma representação difícil, se você tem 4 dimensões ou mais para representar em uma única imagem

Tabela 1 Resultado do modelo de regressão linear múltipla para testar o efeito da quantidade de dinheiro investido na educação básica e proporção de alunos que realizam o teste SAT.

Variável preditora	β	Erro padrão	t	p
Intercepto	1147.10	16.70	68.68	<0.001
Gasto por aluno	11.13	3.26	3.41	0.001
Log (proporção de alunos que fez o teste)	-78.20	4.47	-17.49	<0.001

Nota

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

Esses valores representam os parâmetros do modelo teórico

Tabela 1 Resultado do modelo de regressão linear múltipla para testar o efeito da quantidade de dinheiro investido na educação básica e proporção de alunos que realizam o teste.

Variável preditora	β	Erro padrão	t	p
Intercepto	1147.10	16.70	68.68	<0.001
Gasto por aluno	11.13	3.26	3.41	0.001
Log (proporção de alunos que fez o teste)	-78.20	4.47	-17.49	<0.001

Conclusão (atualizada)

- Sua nova conclusão é que:
- A nota obtida pelos alunos no exame SAT é influenciada simultaneamente pela quantidade de dinheiro que o estado gasta cada aluno (variável Expend) e pelo log da proporção de alunos que realizam o teste (variável LogPctSAT).
- A regressão linear múltipla indicou que:
 - A nota no SAT aumenta com o aumento no gasto com cada aluno
 - A nota no SAT aumenta com a diminuição da proporção de alunos que realizam o teste

Mas qual das duas variáveis é mais importante?

Correlações parciais quadradas

- Correlação parcial (squared partial correlations): mede a correlação entre y e cada variável preditora. Desconsiderando todo o efeito conhecido das outras variáveis predictoras.
- Correlação semiparcial (squared semipartial correlations): proporção de variância de y que é exclusivamente associada a cada variável preditora.

Correlações parciais

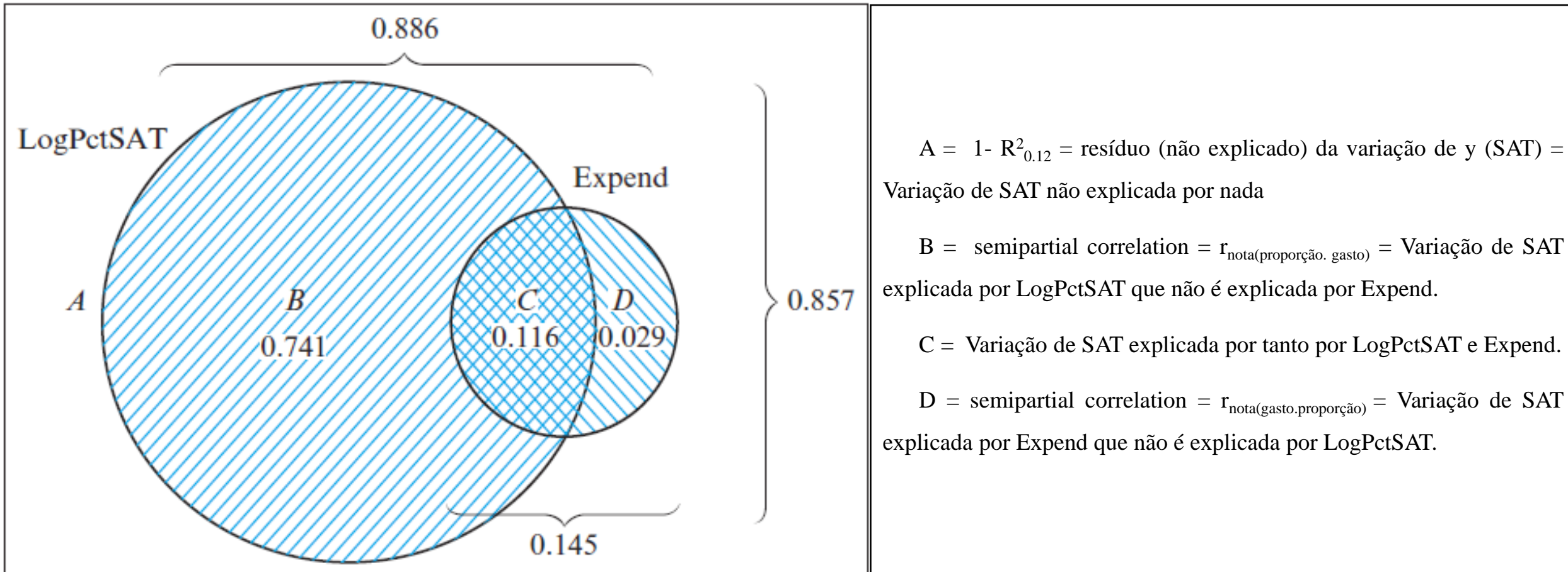


Figure 15.4 Venn diagram illustrating partial and semipartial correlation

Correlações parciais

```
> modelEffectSizes(modelo2)
lm(formula = SAT ~ Expend + LogPctSAT, data = dados)
```

Coefficients

	SSR	df	pEta-sqr	dR-sqr
(Intercept)	3135766.462	1	0.9901	NA
Expend	7725.773	1	0.1983	0.0282
LogPctSAT	203343.849	1	0.8668	0.7413

Sum of squared errors (SSE): 31241.8

Sum of squared total (SST): 274307.7

Correlações parciais

```
> modelEffectSizes(modelo2)
lm(formula = SAT ~ Expend + SAT, data = dados)
```

Coefficients

	SSR	df	pE	sqr	dR	qr
(Intercept)	3135766.462	1	0.9901			NA
Expend	7725.773	1	0.1983	0.0282		
LogPctSAT	203343.849	1	0.8668	0.7413		

Sum of squared errors (SSE): 31241.8

Sum of squared total (SST): 274307.7

squared
partial
correlation

squared
semipartial
correlation

Validação do modelo

Multicolinearidade

- Multicolinearidade = variáveis independentes possuem relações entre elas
- A multicolinearidade vale tanto para variáveis categóricas como variáveis contínuas
- Estimado pelo valor de VIF (variance inflation fator)
 - Valores maiores que 3 são sinal de colinearidade (Zuur et al 2010)

Validação de modelos / VIF

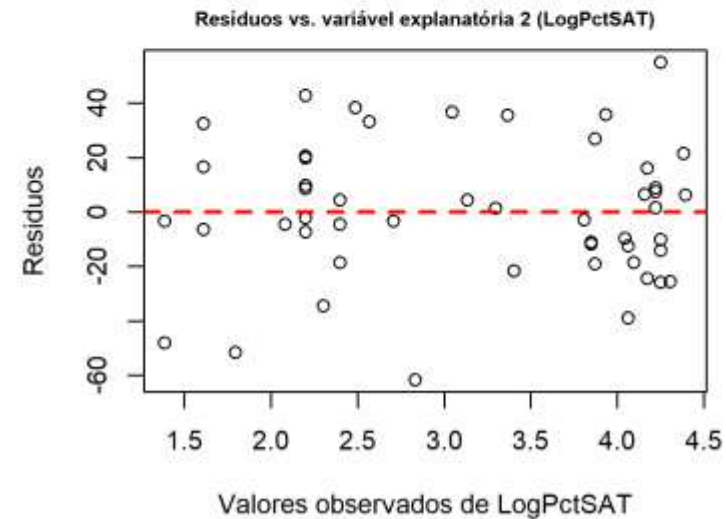
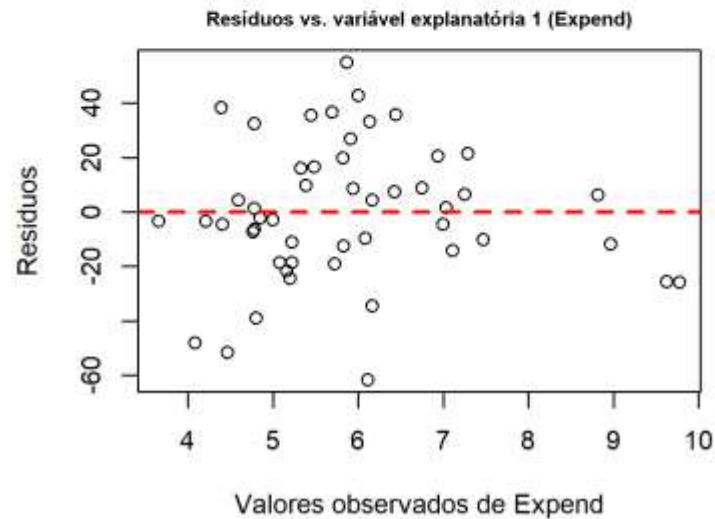
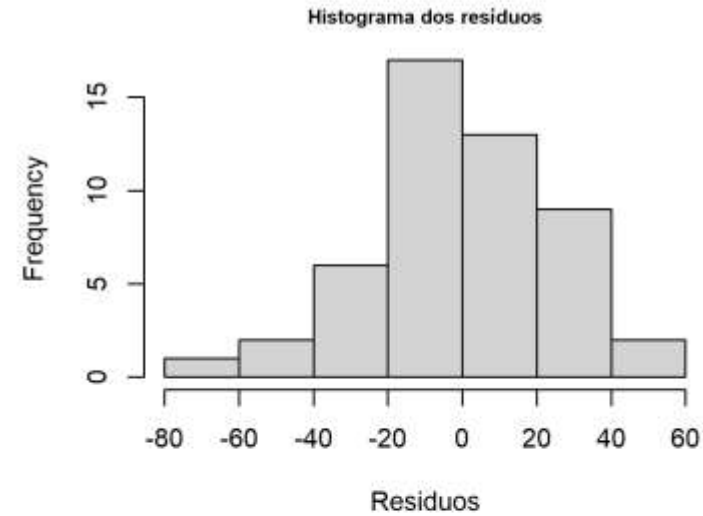
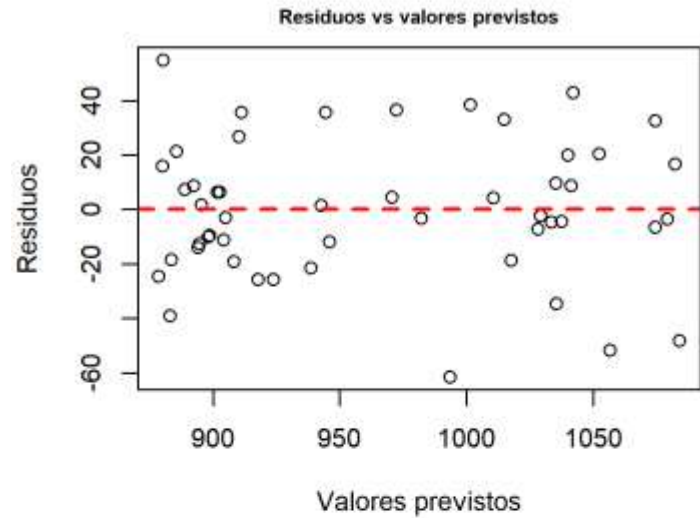
```
> vif(modelo2)
```

Expend	LogPctSAT
1.458884	1.458884

Validação de modelos / Analise de resíduos

- O que são resíduos? R. Diferença entre valores esperados (de acordo com a formula do modelo) e valores observados
 - OBS: Como isso vai aparecer no R
 - Fitted values = valores previstos
 - Residuals = valores previstos – valores observados
- Como analisar os resíduos? Inspeção gráfica (Cap 2, Zuur et al 2009)
 - Fazer um gráfico de resíduos vs valores previstos para avaliar se
 1. A relação é linear
 2. Existe homocedasticidade (pontos estão igualmente distribuídos igualmente em toda a área do gráfico)
 - Avaliar por histograma/boxplot se os resíduos tem distribuição normal
 - Resíduos vs. variável explanatória (para testar independência)

Analise de resíduos



Seleção de modelos

Seleção de modelos

- Quando você está trabalhando com regressão múltipla pode acontecer de:
 - Nem todas as variáveis apresentam efeito significativo
 - Todas as variáveis apresentam efeito significativo, mas um modelo mais simples pode explicar melhor sua variável resposta (Zuur et al 2009)
- Ideia geral: o melhor modelo é aquele que com o menor número de variáveis preditoras, consegue estimar a variável resposta com maior acerto.
- Overfit = quando seu modelo tem tantas variáveis toda variação de y é explicada, mas de tantas variáveis que foram incluídas, seu modelo não explica de fato o fenômeno estudado.

Seleção de modelos

- Como você tem certeza que um modelo matemático é melhor que o outro?
 - Imagine que você quer comparar 4 modelos
 1. $\text{NotaSAT} \sim \text{GastoEducação}$
 2. $\text{NotaSAT} \sim \text{GastoEducação} + \text{ProporçãoTeste}$
 3. $\text{NotaSAT} \sim \text{GastoEducação} + \text{ProporçãoTeste} + \text{Salarios}$
 4. $\text{NotaSAT} \sim 1$
- Obs: modelo 4 diz que nada prevê a nota média de cada estado americano

Seleção de modelos

- Os principais métodos usados para seleção de modelos são
 - Critério de Informação de Akaike (AIC)
 - Critério Bayesiano de Schwarz (BIC)
 - Teste da Razão de Verossimilhança (TRV)

Críticas aos principais métodos de seleção de modelos

- Stepwise é muito comum em artigos de ciências da natureza (Whittingham et al 2006)
- Stepwise infla o erro do tipo I (Whittingham et al 2006; Mundry & Nunn 2009)
- A seleção de modelos por AIC também pode inflar do erro do tipo I (Symonds & Moussalli 2011)

Journal of Animal Ecology 2006
75, 1182–1189

Why do we still use stepwise modelling in ecology and behaviour?

MARK J. WHITTINGHAM, PHILIP A. STEPHENS*, RICHARD B. BRADBURY† and ROBERT P. FRECKLETON‡

*Division of Biology, School of Biology and Psychology, Ridley Building, University of Newcastle, Newcastle Upon Tyne, NE1 7RU, UK; *Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK; †Royal Society for the Protection of Birds, The Lodge, Sandy, Bedfordshire, SG19 2DL, UK; and ‡Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK*

VOL. 173, NO. 1 THE AMERICAN NATURALIST JANUARY 2009

Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution

Roger Mundry^{1,*} and Charles L. Nunn^{1,2,3,†}

Behav Ecol Sociobiol (2011) 65:13–21
DOI 10.1007/s00265-010-1037-6

REVIEW

A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion

Matthew R. E. Symonds • Adnan Moussalli

Seleção de modelos: o que fazer?

- Opção I: não fazer seleção de modelos
 - Pense em um modelo que faz sentido para você
 - Testar o efeito das variáveis preditoras uma única vez
 - Aceite como evidência de efeito significativo apenas as variáveis com $p < 0.05$ no seu modelo
- Opção II: aceitar que seu estudo pode ter o erro tipo I aumentado
 - Reconheça a possibilidade do erro tipo I ter sido aumentado
 - Faça a seleção de modelos por algum dos métodos indicados

Opção I: não fazer seleção de modelos

Exemplo possibilidade I (não fazer seleção de modelos)

- Modelo completo (full model)

$$\text{NotaSAT} \sim \text{GastoEducação} + \text{ProporçãoTeste} + \text{Salarios}$$

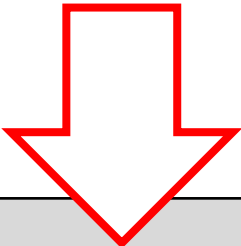
Tabela X Resultado do modelo de regressão linear múltipla

Variável preditora	β	Erro padrão	t	p
Intercepto	1133.1	22.73	49.86	<0.001
Gasto por aluno	7.10	5.50	1.29	0.20
Log (proporção de alunos que fez o teste)	-79.5	4.70	-16.90	<0.001
Salario dos professores	1.20	1.32	0.91	0.37

Exemplo possibilidade I (não fazer seleção de modelos)

- Identifique os preditores que apresentaram efeito significativo e interprete o resultado:
- Conclusão: A nota média de alunos no exame SAT é prevista apenas pela proporção de alunos que realizam o teste

Tabela X Resultado do modelo de regressão linear múltipla



Variável preditora	β	Erro padrão	t	p
Intercepto	1133.1	22.73	49.86	<0.001
Gasto por aluno	7.10	5.50	1.29	0.20
Log (proporção de alunos que fez o teste)	-79.5	4.70	-16.90	<0.001
Salario dos professores	1.20	1.32	0.91	0.37

Opção II: fazer seleção de modelos

Possibilidades

- Criar modelos contendo as diferentes possibilidades de variáveis preditoras, e testar qual é o modelo que melhor se ajusta aos dados por meio de teste de AIC
- Criar modelos contendo todas as variáveis preditoras, e reduzir a complexidade do modelo pelo método de *stepwise model selection* (Zuur et al 2009)

AIC e BIC

- Modelo com menor AIC ou BIC é considerado o de melhor ajuste

```
> AIC(modelo1,modelo2,modelo3,modelo4)
```

	df	AIC
modelo1	3	570.5715
modelo2	4	471.7683
modelo3	5	472.8765
modelo4	2	576.3930

Modelo 2 é o melhor

```
> BIC(modelo1,modelo2,modelo3,modelo4)
```

	df	BIC
modelo1	3	576.3076
modelo2	4	479.4164
modelo3	5	482.4366
modelo4	2	580.2170

Modelo 2 é o melhor

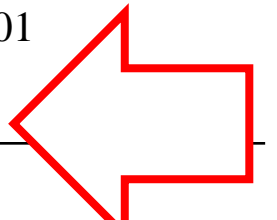
OBS: Modelo 2: NotaSAT ~ GastoEducação + ProporçãoTeste

Stepwise

- Observe o efeito de cada variável
- Caso haja variável com $p > 0,05$
- Refaça a análise sem a variável que tinha o maior valor de p
- Repita isso até que o modelo tenha apenas variáveis com efeito significativo

Tabela X Resultado do modelo de regressão linear múltipla

Variável preditora	β	Erro padrão	t	p
Intercepto	1133.1	22.73	49.86	<0.001
Gasto por aluno	7.10	5.50	1.29	0.20
Log (proporção de alunos que fez o teste)	-79.5	4.70	-16.90	<0.001
Salario dos professores	1.20	1.32	0.91	0.37



Stepwise

- Observe o efeito de cada variável
- Caso haja variável com $p > 0,05$
- Refaça a análise sem a variável que tinha o maior valor de p
- Repita isso até que o modelo tenha apenas variáveis com efeito significativo

Tabela X Resultado do modelo de regressão linear múltipla

Variável preditora	β	Erro padrão	t	p
Intercepto	1147.10	16.70	68.68	<0.001
Gasto por aluno	11.13	3.26	3.41	0.001
Log (proporção de alunos que fez o teste)	-78.20	4.47	-17.49	<0.001

Todas apresentam efeito significativo

Forward model selection

- Análoga a stepwise, esse método consiste em você adicionar as variáveis preditoras uma a uma, e comparar os valores de AIC.
- O modelo final será aquele que a adição de novas variáveis não diminuam o valor de AIC

Forward model selection

Start: AIC=432.5
SAT ~ 1

Observações	
Passo	1
Formula do modelo	SAT ~1
AIC modelo atual	432,5
AIC do modelo com o termo LogPctSat	336,92

	Df	Sum of Sq	RSS	AIC
+ LogPctSAT	1	235340	38968	336.92
+ Salary	1	53078	221230	423.75
+ Expend	1	39722	234586	426.68
<none>			274308	432.50

Identifique se algum termo reduz o AIC
Se sim, qual termo que reduz mais significativamente o AIC

Forward model selection

Step: AIC=336.92
SAT ~ LogPctSAT

	Df	Sum of Sq	RSS	AIC
+ Expend	1	7725.8	31242	327.87
+ Salary	1	7165.8	31802	328.76
<none>			38968	336.92

Observações

Passo	1
Formula do modelo	SAT ~LogPcSAT
AIC modelo atual	336,92
AIC do modelo com o termo LogPctSat	327,92

Identifique se algum termo reduz o AIC
Se sim, qual termo que reduz mais significativamente o AIC

Forward model selection

Observações	
Passo	2
Formula do modelo	SAT ~LogPcSAT + Expend
AIC modelo atual	327,92
AIC do modelo novo modelo	Nenhum outro termo diminuía o valor de AIC

Step: AIC=327.87
SAT ~ LogPctSAT + Expend

	Df	Sum of Sq	RSS	AIC
<none>			31242	327.87
+ salary	1	552.29	30690	328.98

Identifique se algum termo reduz o AIC
Como aqui isso não foi observado, paramos com esse modelo

Forward model selection

- Conclusão, o melhor modelo é: $SAT \sim \text{LogPctSAT} + \text{Expend}$

call:

```
lm(formula = SAT ~ LogPctSAT + Expend, data = dados)
```

Coefficients:

(Intercept)	LogPctSAT	Expend
1147.10	-78.20	11.13

Prática

A prática de hoje será refazer os
exemplo apresentados na aula