CE225 - Modelos Lineares Generalizados

Cesar Augusto Taconeli

01 de agosto, 2017

Aula 1 - Introdução

Uma breve reflexão...

George Box

All models are wrong but some are useful

Richard Feynman

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.

John W. Tukey

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.

Modelos Lineares

- Exemplos de modelos lineares:
  - Modelos de regressão linear;
  - Modelos de análise de variância;
  - Modelos de análise de covariância.
- Nesta disciplina, frequentemente vamos usar o termo **regressão** de forma genérica, contemplando toda a classe de modelos lineares (generalizados).

Modelos Lineares

- Modelos lineares descrevem a relação entre uma variável aleatória (resposta) e um conjunto de variáveis (fatores) explicativas.

- Algumas restrições se aplicam aos modelos lineares:
  - A relação entre as variáveis (reposta e explicativas) é descrita por um conjunto de parâmetros, por meio de uma função linear;
  - Condicional aos valores das variáveis explicativas, as respostas são independentes, tem distribuição Normal e igual variância.

- Embora válidas em muitos casos, tais suposições nem sempre são satisfeitas, tornando necessária a utilização de métodos mais flexíveis.

Modelos Lineares Generalizados

- **Origem:** Nelder e Wedderburn (1972): "Generalized Linear Models", publicado em *Journal of the Royal Statistical Society*;
- Extensão dos modelos lineares, incorporando, sob uma teoria unificada, diversos outros modelos propostos até então;
- Tais modelos permitem contemplar, num contexto de análise de regressão, variáveis respostas pertencentes à **família exponencial** de distribuições;
- Como casos particulares da família exponencial temos as distribuições binomial, Poisson, Normal, Gama e Normal Inversa, dentre outras.

Ilustração - alguns problemas abordados em MLG

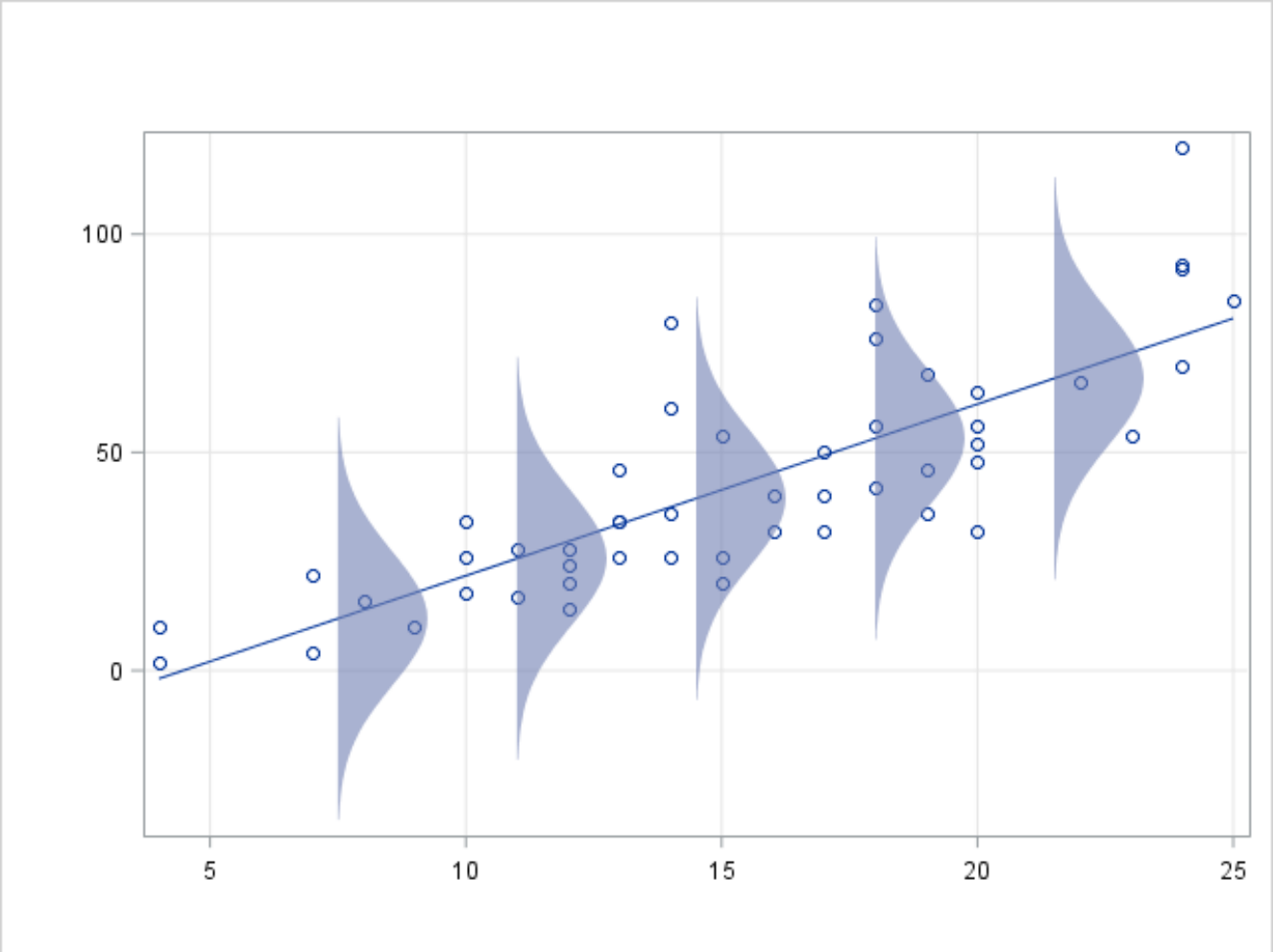

Figura 1: Regressão com erros normais - I

Ilustração - alguns problemas abordados em MLG

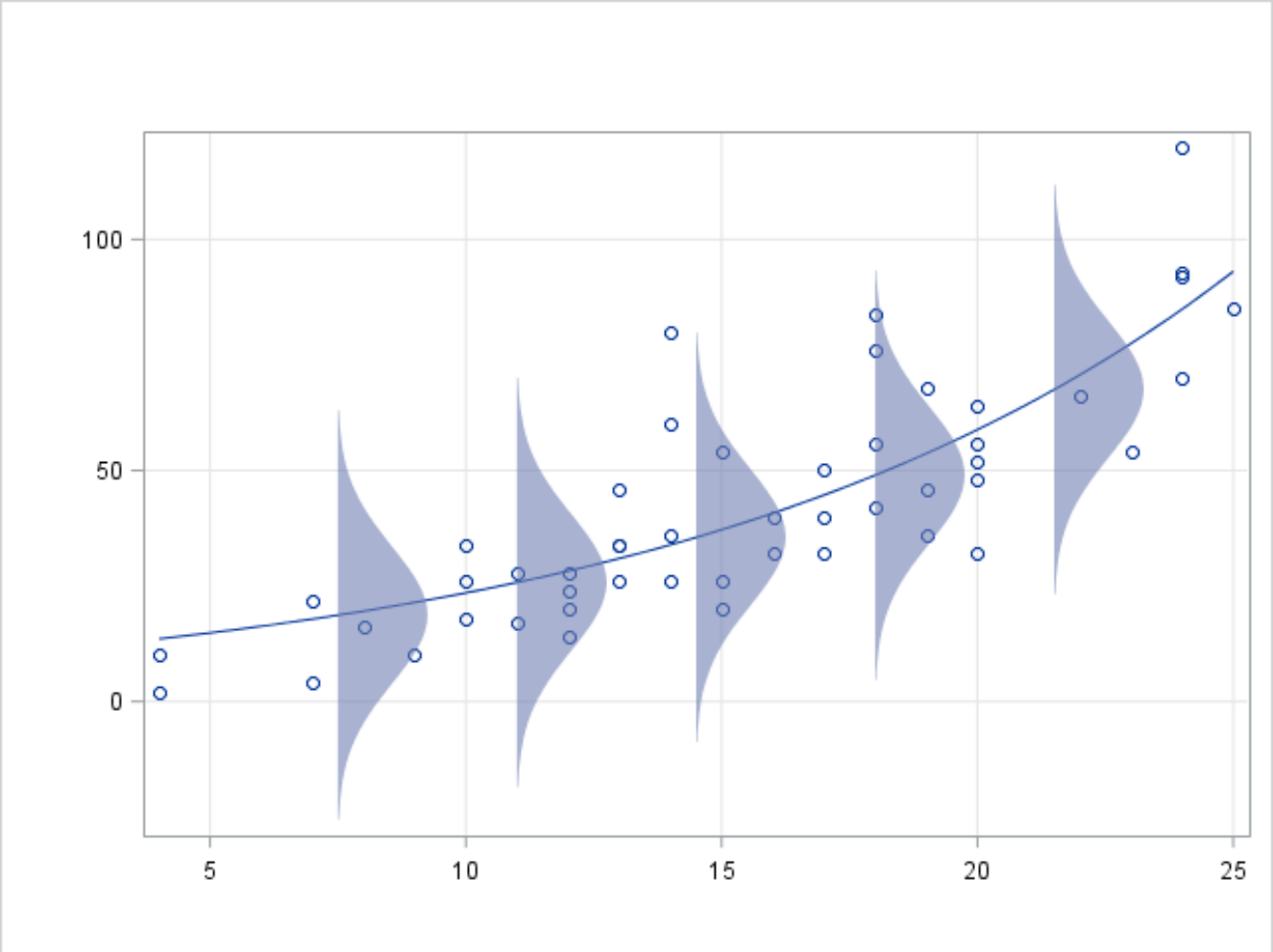

Figura 2: Regressão com erros normais - II

Ilustração - alguns problemas abordados em MLG

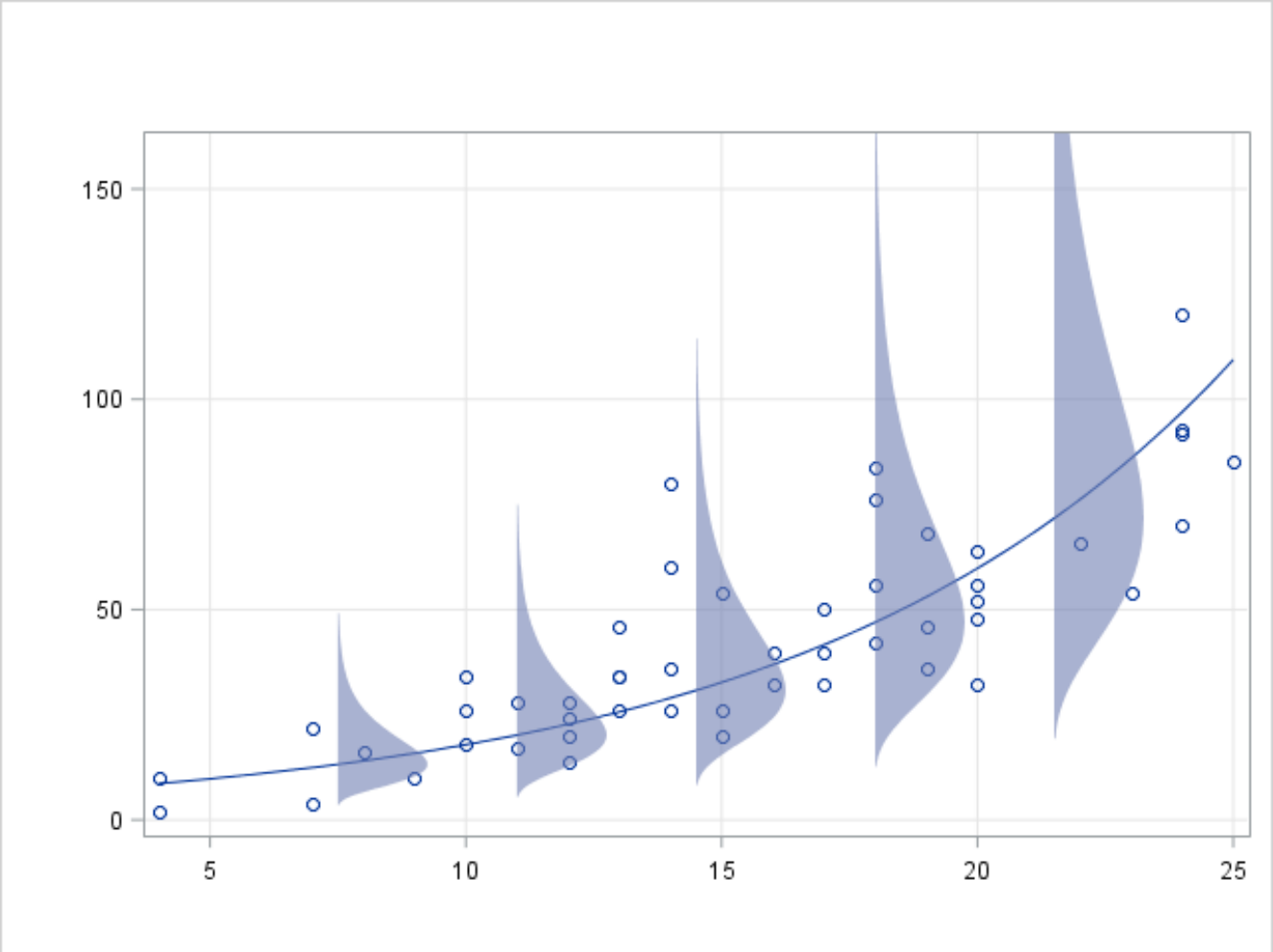

Figura 3: Regressão para dados contínuos assimétricos

Ilustração - alguns problemas abordados em MLG

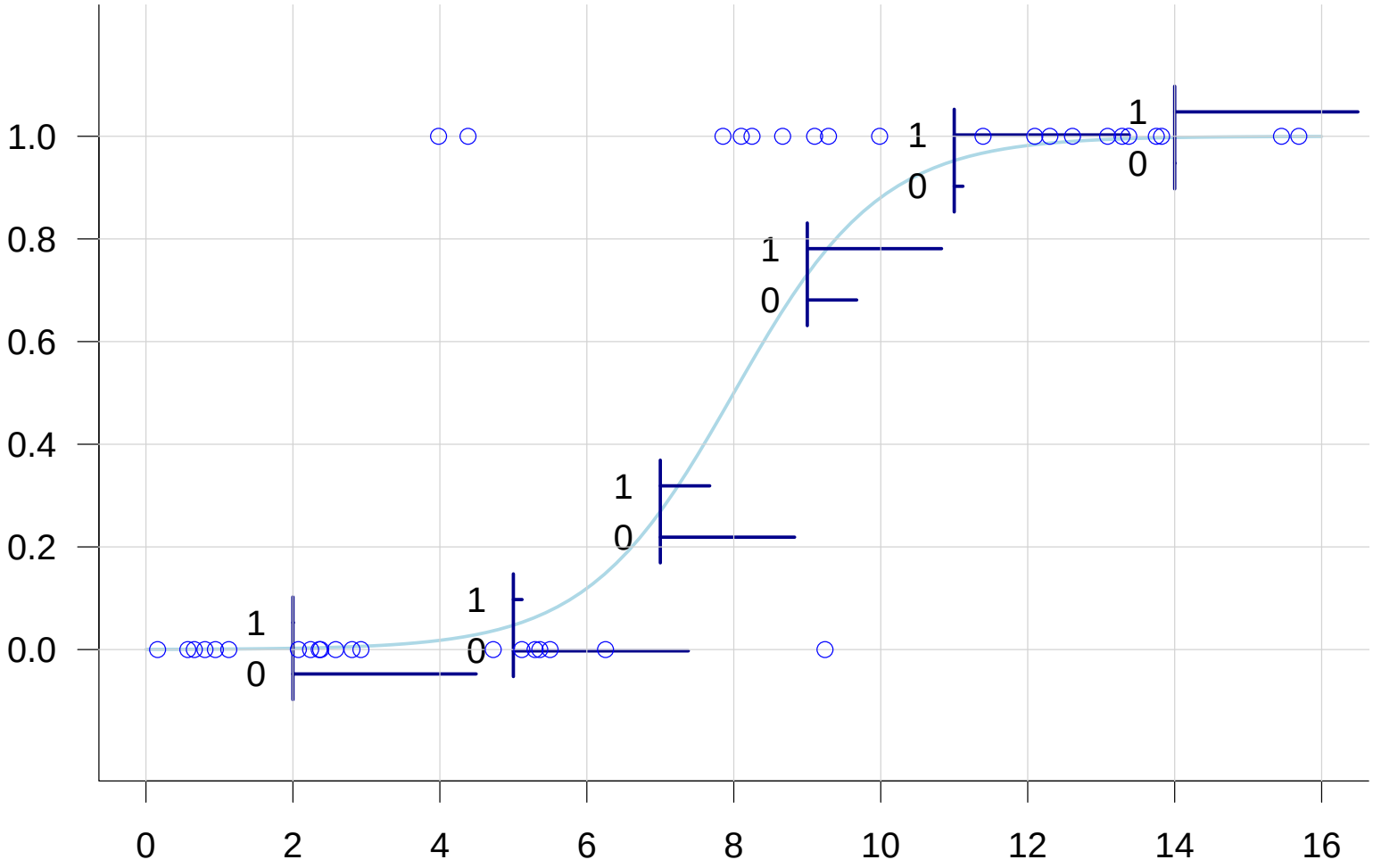

Figura 4: Regressão para dados binários

Ilustração - alguns problemas abordados em MLG

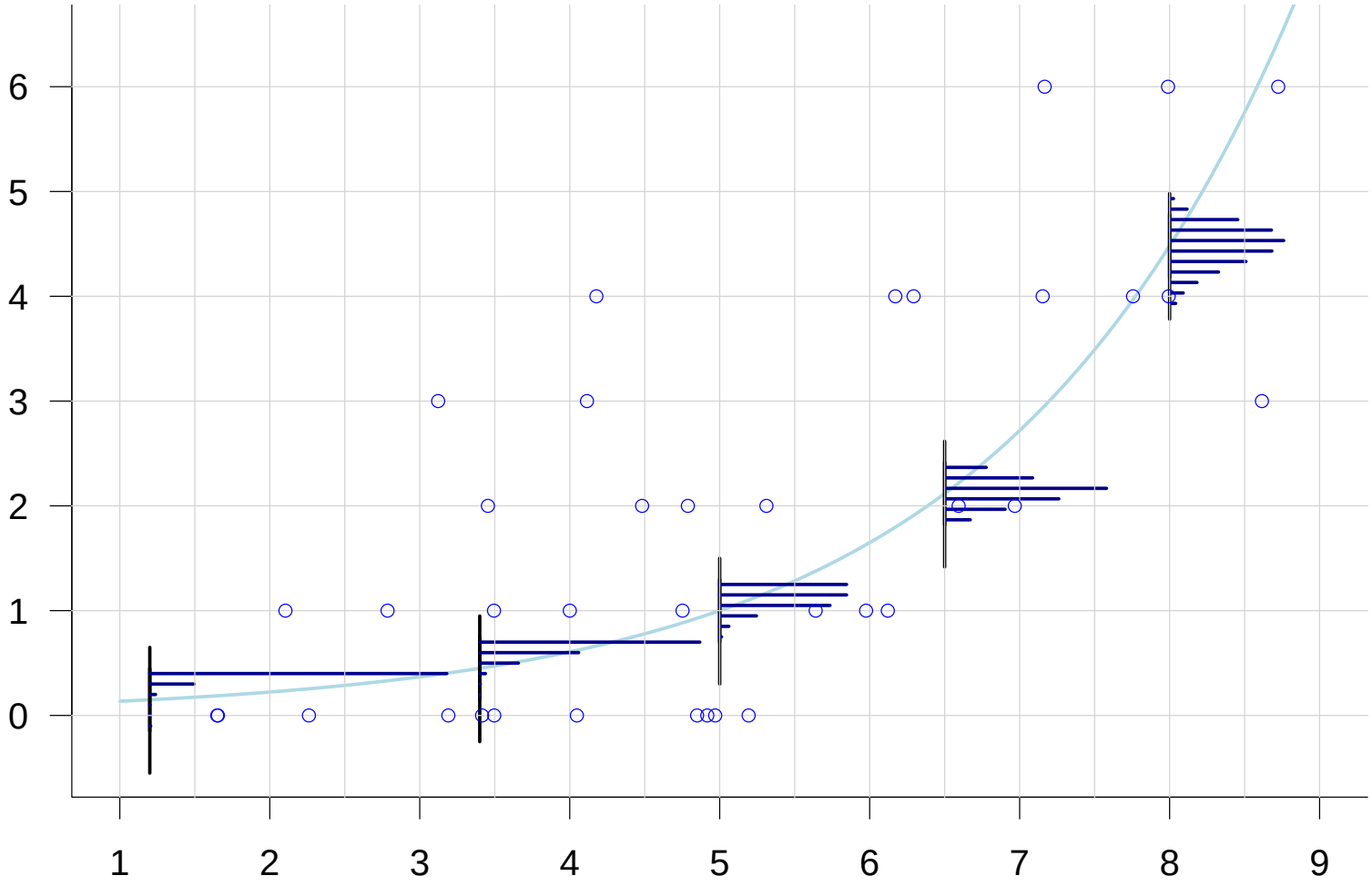

Figura 5: Regressão para dados de contagens

Exemplos de motivação

Exemplo 1 - Resistência de fibra sintética

Exemplo 1

Dados de um experimento planejado com o objetivo de avaliar a resistência de fibra sintética usada na fabricação de camisas. Foram considerados tecidos com diferentes quantidades de algodão em sua composição.

- **Variável resposta:** Resistência da fibra ($libras/pol^2$);
- **Variável explicativa:** Porcentagem de algodão no tecido, fator com cinco níveis (15, 20, 25, 30 e 35%).

Exemplo 1 - Resistência de fibra sintética

Tabela 1: Resistência (em $libras/pol^2$) das amostras de tecido.

% Algodão	Resistência do tecido				
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

Exemplo 1 - Resistência de fibra sintética

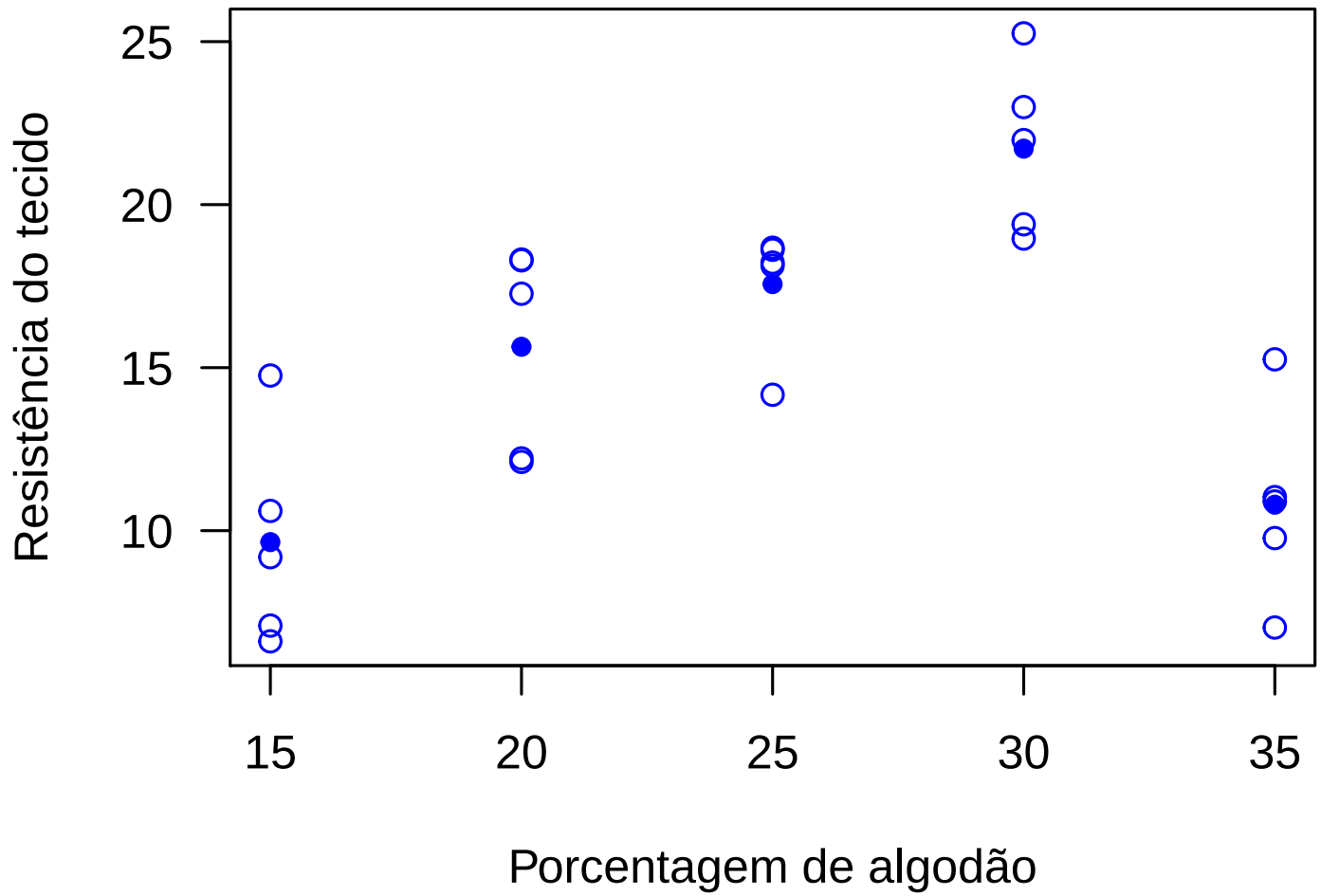

Figura 6: Gráfico de dispersão para as resistências das fibras sob cinco diferentes porcentagens de algodão

Exemplo 1 - Resistência de fibra sintética

- **Objetivos da análise:**

  * Analisar o efeito da porcentagem de algodão na resistência da fibra sintética;

  * Estimar a porcentagem ótima de algodão (aquela que proporciona máxima resistência).

Exemplo 2 - Mortalidade da praga do algodão

Exemplo 2

Dados de um experimento planejado com o objetivo de avaliar a mortalidade de insetos submetidos a doses crescentes de cipermetrina. Vinte insetos machos e 20 fêmeas foram submetidos a cada dose. Após 72 horas de experimento, foram contados os insetos mortos.

- **Variável resposta:** Número de insetos mortos;
- **Variáveis explicativas:**
  - Dose de cipermetrina: 1, 2, 4, 8, 16, 32 u.m.;
  - Sexo (Macho ou Fêmea).

Exemplo 2 - Mortalidade da praga do algodão

Tabela 2: Números de insetos mortos para as diferentes doses de cipermetrina

Dose	Log2(Dose)	Machos	Fêmeas
1	0	1	0
2	1	4	2
4	2	9	6
8	3	13	10
16	4	18	12
32	5	20	16

Exemplo 2 - Mortalidade da praga do algodão

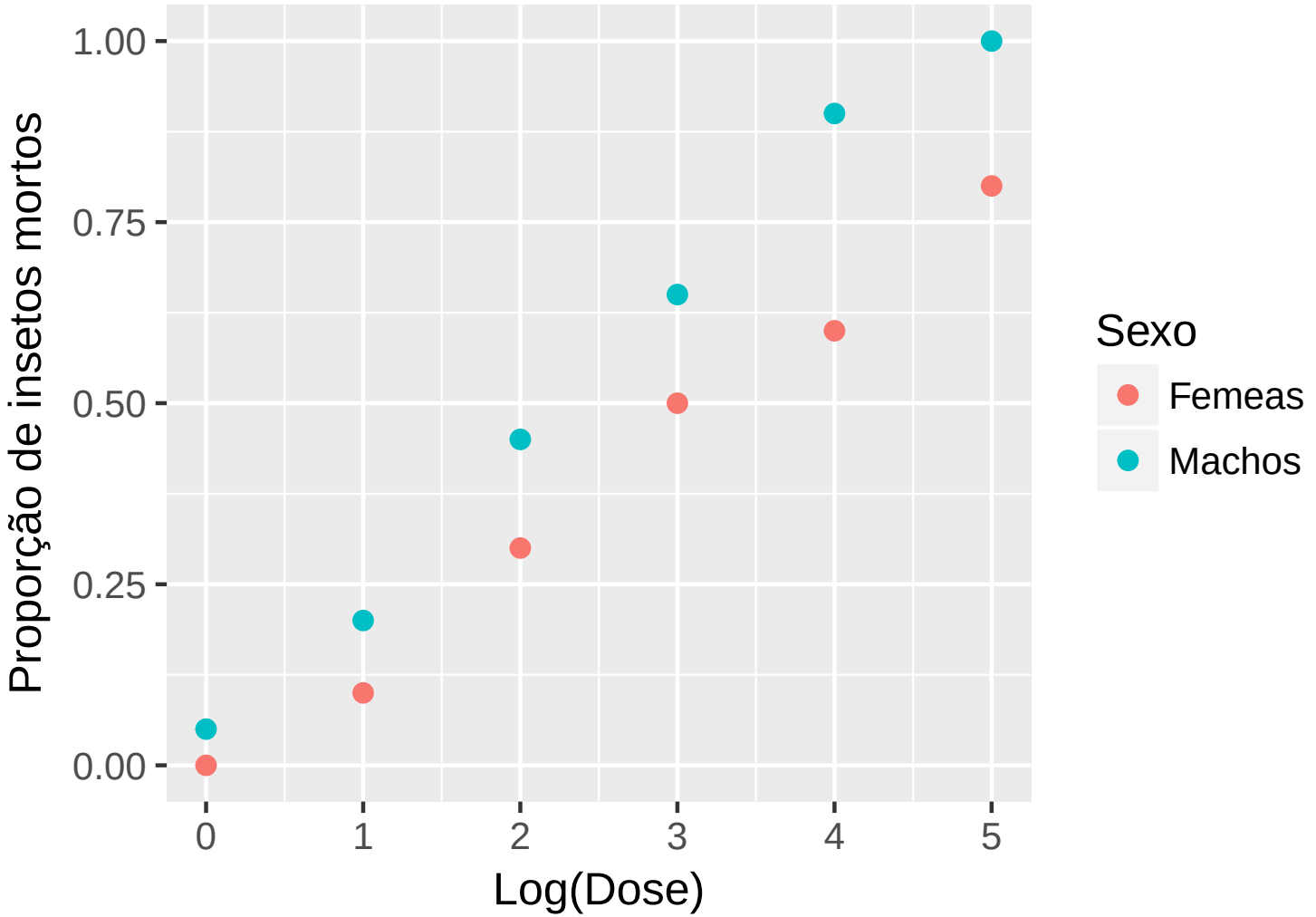

Figura 7: Proporção de insetos mortos segundo sexo e dose de inseticida

Exemplo 2 - Mortalidade da praga do algodão

- **Objetivos da análise:**
  - Descrever (modelar) a variação na mortalidade de insetos segundo a dose aplicada de inseticida;
  - Comparar as curvas de mortalidade de insetos machos e fêmeas;
  - Estimar doses efetivas (letais), que matam determinada proporção de insetos.

Exemplo 3 - Diagnóstico de diabetes em mulheres indígenas

Exemplo 3

Diagnóstico de diabetes e outras variáveis clínicas avaliadas em uma amostra de mulheres adultas indígenas de uma comunidade próxima a Phoenix, Arizona. A amostra contém os registros completos de 532 habitantes.

Exemplo 3 - Diagnóstico de diabetes em mulheres indígenas

- **Variável resposta:**
  - **Diabetes:** Diagnóstico de diabetes de acordo com o teste de glicemia em jejum (0 - Negativo; 1 - Positivo);
- **Variáveis explicativas:**
  - **Gest:** Número de gestações;
  - **GlicOral:** Concentração de glicose no teste oral de tolerância à glicose;
  - **Pressao:** Pressão arterial diastólica (em mmHg);
  - **Prega:** Espessura da prega tricipital (mm);
  - **IMC:** Índice de massa corporal ($peso/altura^2$);
  - **Pedigree:** Índice referente ao histórico de diabetes na família;
  - **Idade:** em anos.

Exemplo 3 - Diagnóstico de diabetes em mulheres indígenas

Tabela 3: Primeiras linhas da base

	Gest	GlicOral	Pressao	Prega	IMC	Pedigree	Idade	Diabetes
1	6	148	72	35	33.6	0.627	50	Sim
2	1	85	66	29	26.6	0.351	31	Não
4	1	89	66	23	28.1	0.167	21	Não
5	0	137	40	35	43.1	2.288	33	Sim
7	3	78	50	32	31.0	0.248	26	Sim
9	2	197	70	45	30.5	0.158	53	Sim
14	1	189	60	23	30.1	0.398	59	Sim
15	5	166	72	19	25.8	0.587	51	Sim
17	0	118	84	47	45.8	0.551	31	Sim
19	1	103	30	38	43.3	0.183	33	Não

Exemplo 3 - Diagnóstico de diabetes em mulheres indígenas

Tabela 4: Resumo - dados sobre diabetes

Gest	GlicOral	Pressao	Prega
Min. : 0.000	Min. : 56.00	Min. : 24.00	Min. : 7.00
1st Qu.: 1.000	1st Qu.: 98.75	1st Qu.: 64.00	1st Qu.:22.00
Median : 2.000	Median :115.00	Median : 72.00	Median :29.00
Mean : 3.517	Mean :121.03	Mean : 71.51	Mean :29.18
3rd Qu.: 5.000	3rd Qu.:141.25	3rd Qu.: 80.00	3rd Qu.:36.00
Max. :17.000	Max. :199.00	Max. :110.00	Max. :99.00

Tabela 5: Resumo - dados sobre diabetes (cont)

IMC	Pedigree	Idade	Diabetes
Min. :18.20	Min. :0.0850	Min. :21.00	Não:355
1st Qu.:27.88	1st Qu.:0.2587	1st Qu.:23.00	Sim:177
Median :32.80	Median :0.4160	Median :28.00	
Mean :32.89	Mean :0.5030	Mean :31.61	
3rd Qu.:36.90	3rd Qu.:0.6585	3rd Qu.:38.00	
Max. :67.10	Max. :2.4200	Max. :81.00	

Exemplo 3 - Diagnóstico de diabetes em mulheres indígenas

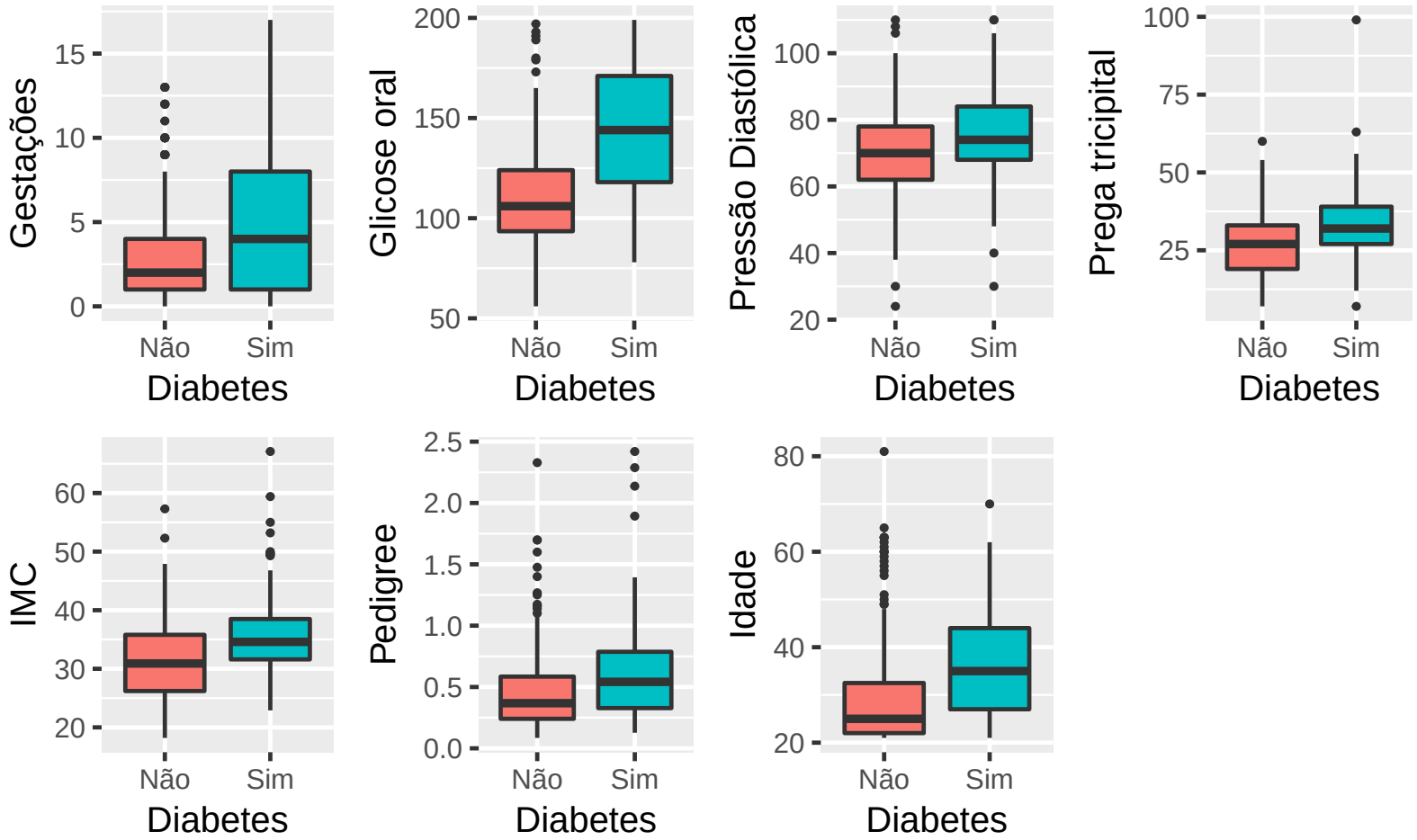

Figura 8: Distribuição das variáveis explicativas segundo o diagnóstico de diabetes

Exemplo 3 - Diagnóstico de diabetes em mulheres indígenas

- **Objetivos da análise:**

- Determinar um modelo preditivo para o diagnóstico de diabetes, como alternativa ao teste de glicemia em jejum.

- Identificar fatores de risco associados à diabetes.

Exemplo 4 - Acasalamento de elefantes

Exemplo 4

Dados referentes ao número de acasalamentos bem sucedidos e idades de 41 elefantes machos de uma população africana.

- **Variável resposta:**
  - **Matings:** Número de acasalamentos bem sucedidos;
- **Variável explicativa:**
  - **Age:** Idade (em anos).

Exemplo 4 - Acasalamento de elefantes

Tabela 6: Dez linhas da base selecionadas ao acaso para visualização

Age	Matings
44	3
28	1
33	3
33	4
48	2
34	3
29	2
27	0
28	1
36	5

Exemplo 4 - Acasalamento de elefantes

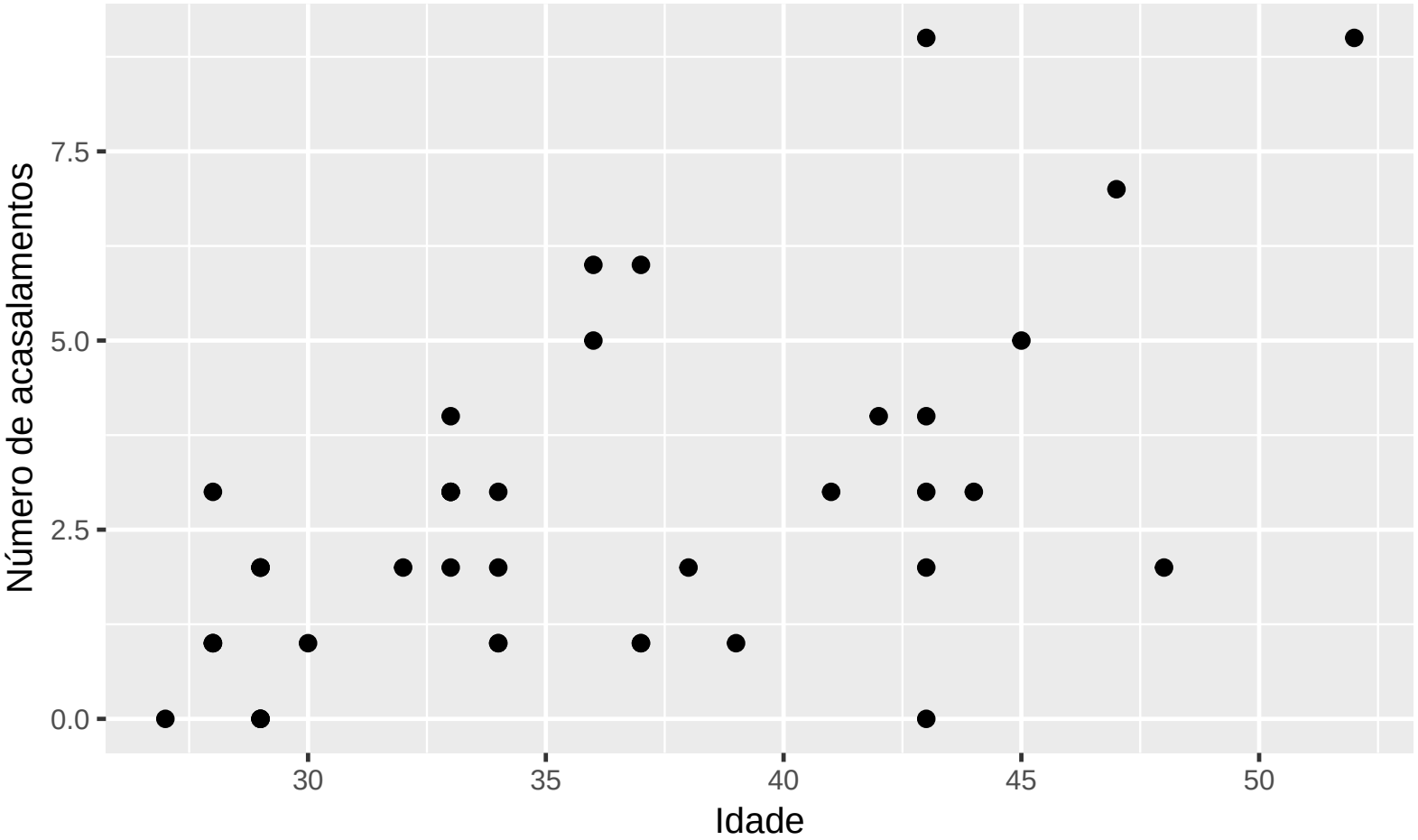

Figura 9: Número de acasalamentos versus idade.

Exemplo 4 - Acasalamento de elefantes

- **Objetivos da análise:**

- Analisar se há predominância de animais mais velhos na incidência de acasalamentos (o que pode induzir maior longevidade da espécie, pela transmissão da carga genética).
- Estimar a variação na taxa de acasalamentos bem sucedidos conforme a idade.

Exemplo 5 - Infecções de ouvido em soldados norte-americanos

Exemplo 5

Dados referentes à incidência de infecções de ouvido em uma amostra de 287 soldados norte-americanos durante o ano de 1990.

- **Variável resposta:**
  - **ninfec:** Número de episódios de infecção (auto-declarado);
- **Variáveis explicativas:**
  - **habito:** Frequência com que costuma nadar (ocasional ou frequente);
  - **local:** Local em que costuma nadar (praia ou piscina);
  - **idade:** Categorizada em três faixas (15-19, 20-24 e 25-29);
  - **sexo:** F: feminino; M: masculino.

Exemplo 5 - Infecções de ouvido em soldados norte-americanos

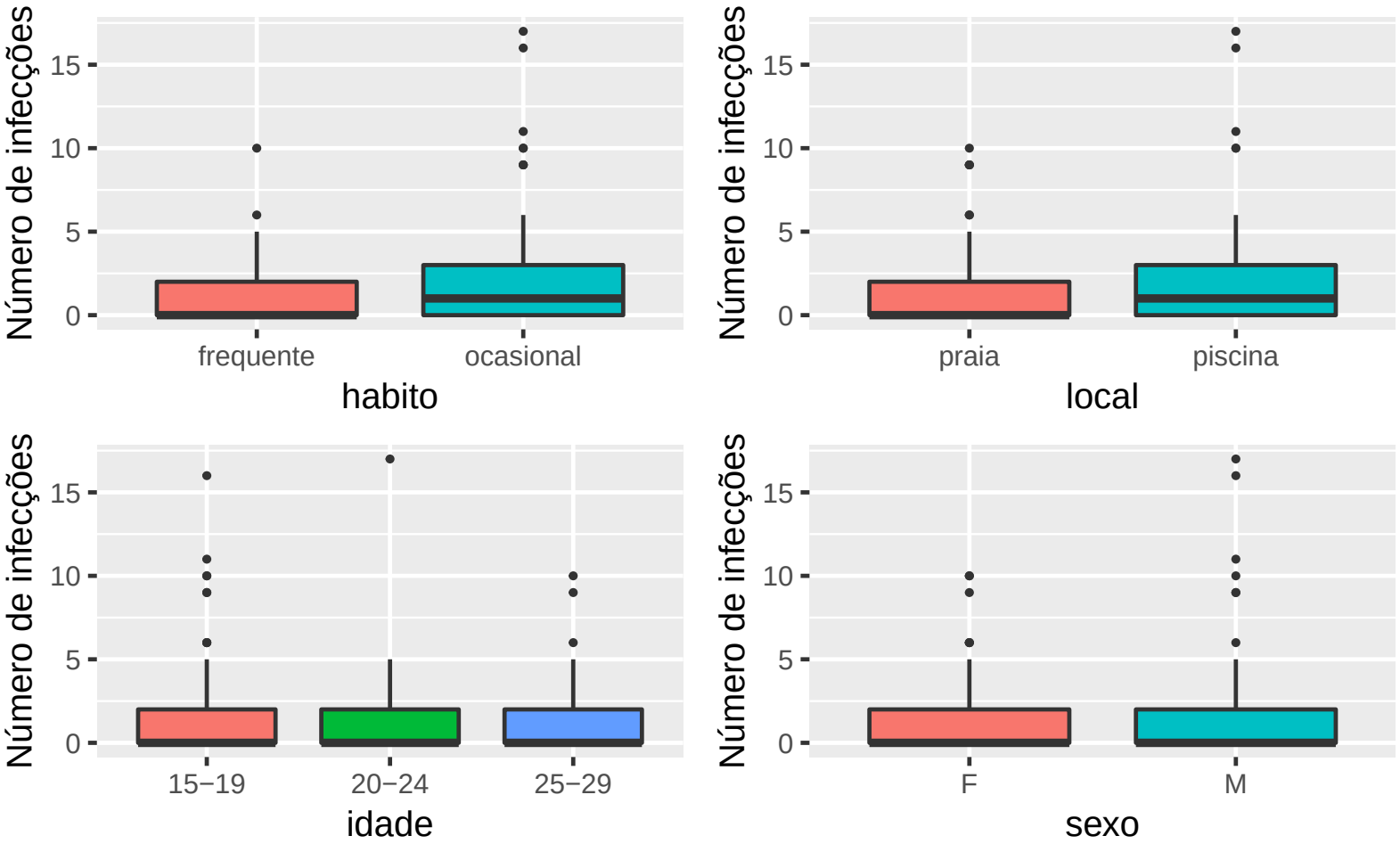

Figura 10: Distribuição das frequências de episódios de infecção no ouvido

Exemplo 5 - Infecções de ouvido em soldados norte-americanos

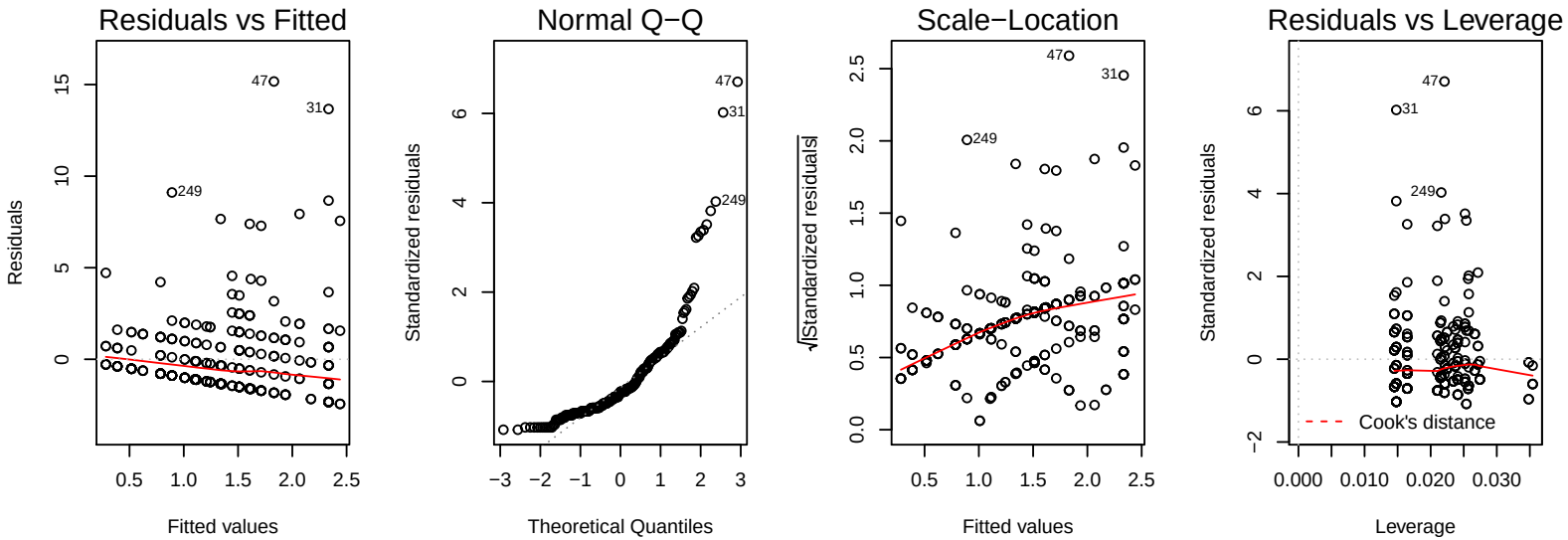

Figura 11: Gráficos de resíduos para o ajuste de um modelo linear

Exemplo 5 - Infecções de ouvido em soldados norte-americanos

- **Objetivos da análise:**

- Verificar se há associação entre a frequência e o local onde os soldados costumam nadar e a incidência de infecção nos ouvidos.

- Identificar perfis de soldados mais propensos a apresentar infecção.

Exemplo 6 - Sinistros em apólices de seguros de automóveis

Exemplo 6

Dados de 4624 apólices de seguros de automóveis que registraram sinistro no período de um ano, entre 2004 e 2005, para uma seguradora.

- **Variável resposta:**
  - **claimcst0:** Valor (somado) dos sinistros apresentados no período.
- **Variáveis explicativas:**
  - **veh_value:** Valor do veículo (em 10.000 dólares);
  - **veh_body:** Tipo de veículo (12 categorias);
  - **veh_age:** Idade do veículo (em quatro níveis - 1, 2, 3 ou 4, dos mais novos aos mais antigos);
  - **gender:** Sexo do motorista principal (F: feminino; M: masculino);
  - **area:** Área da residência do motorista (seis áreas - A, B, C, D, E e F);
  - **agecat:**Idade do motorista (em quatro níveis - 1, 2, 3, 4, 5 ou 6, dos mais novos aos mais velhos).

Exemplo 6 - Sinistros em apólices de seguros de automóveis

Tabela 7: Dez primeiras linhas da base

	claimcst0	veh_value	veh_body	veh_age	gender	area	agecat
15	0.0669510	1.66	SEDAN	3	M	B	6
17	0.0806610	1.51	SEDAN	3	F	F	4
18	0.0401805	0.76	HBACK	3	M	C	4
41	0.1811710	1.89	STNWG	3	M	F	2
65	0.5434440	4.06	STNWG	2	M	F	3
66	0.0865790	1.39	HBACK	3	F	A	4
96	0.1105770	2.66	STNWG	1	F	F	5
99	0.0200000	0.50	HBACK	4	F	A	5
116	0.0739230	1.16	STNWG	4	F	B	2
125	0.3230600	3.56	MCARA	3	M	F	4

Exemplo 6 - Sinistros em apólices de seguros de automóveis

Tabela 8: Resumo - dados sobre valores de sinistros

claimcst0	veh_value	veh_body	veh_age	gender	area	agecat
Min. :0.02000	Min. : 0.000	SEDAN :1476	1: 825	F:2648	A:1085	1: 496
1st Qu.:0.03538	1st Qu.: 1.100	HBACK :1264	2:1259	M:1976	B: 965	2: 932
Median :0.07616	Median : 1.570	STNWG :1173	3:1362		C:1412	3:1113
Mean :0.20144	Mean : 1.859	UTE : 260	4:1178		D: 496	4:1104
3rd Qu.:0.20914	3rd Qu.: 2.310	HDTOP : 130			E: 386	5: 614
Max. :5.59221	Max. :13.900	TRUCK : 120			F: 280	6: 365
		(Other): 201				

Exemplo 6 - Sinistros em apólices de seguros de automóveis

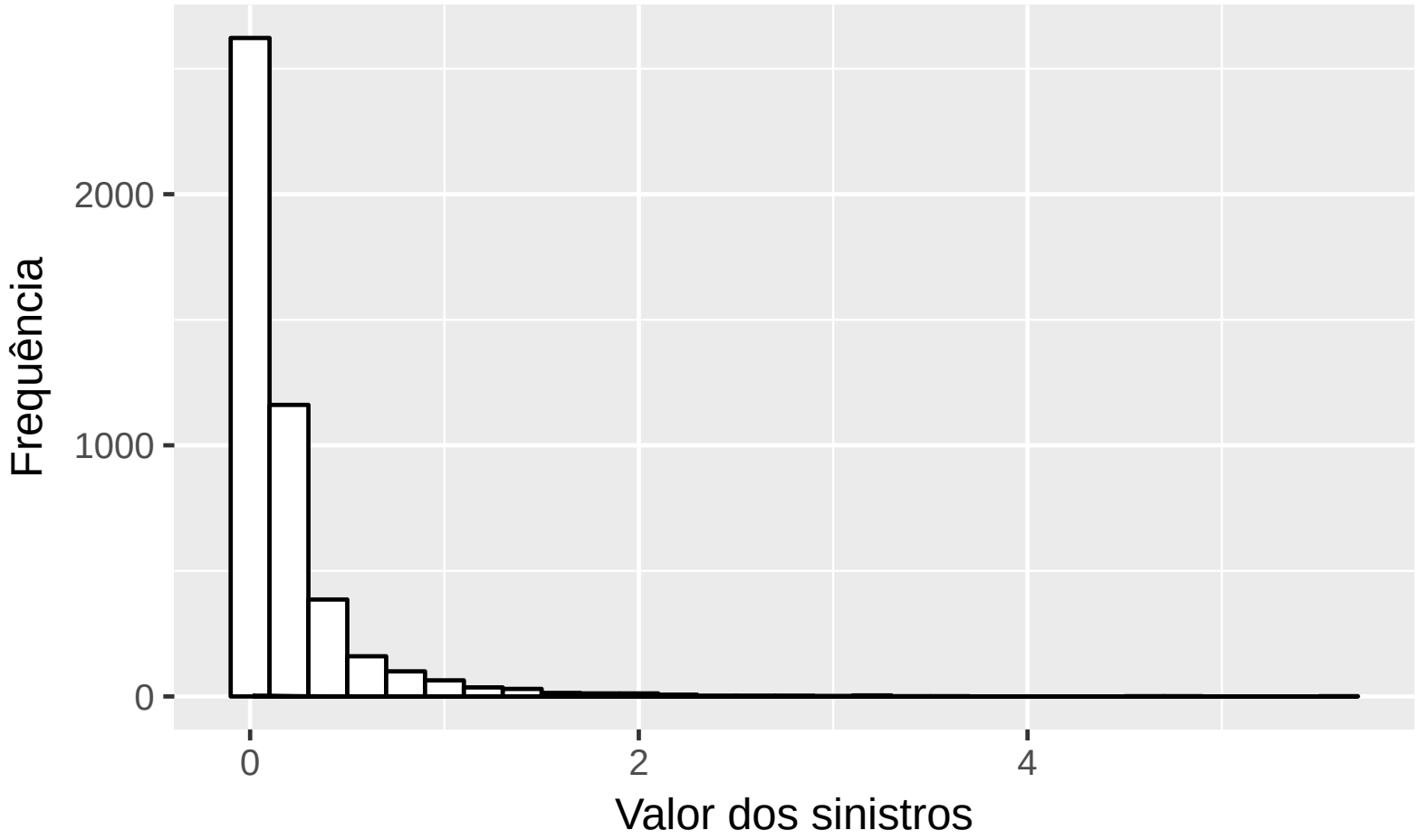

Figura 12: Distribuição de frequências - seguros de automóveis

Exemplo 6 - Sinistros em apólices de seguros de automóveis

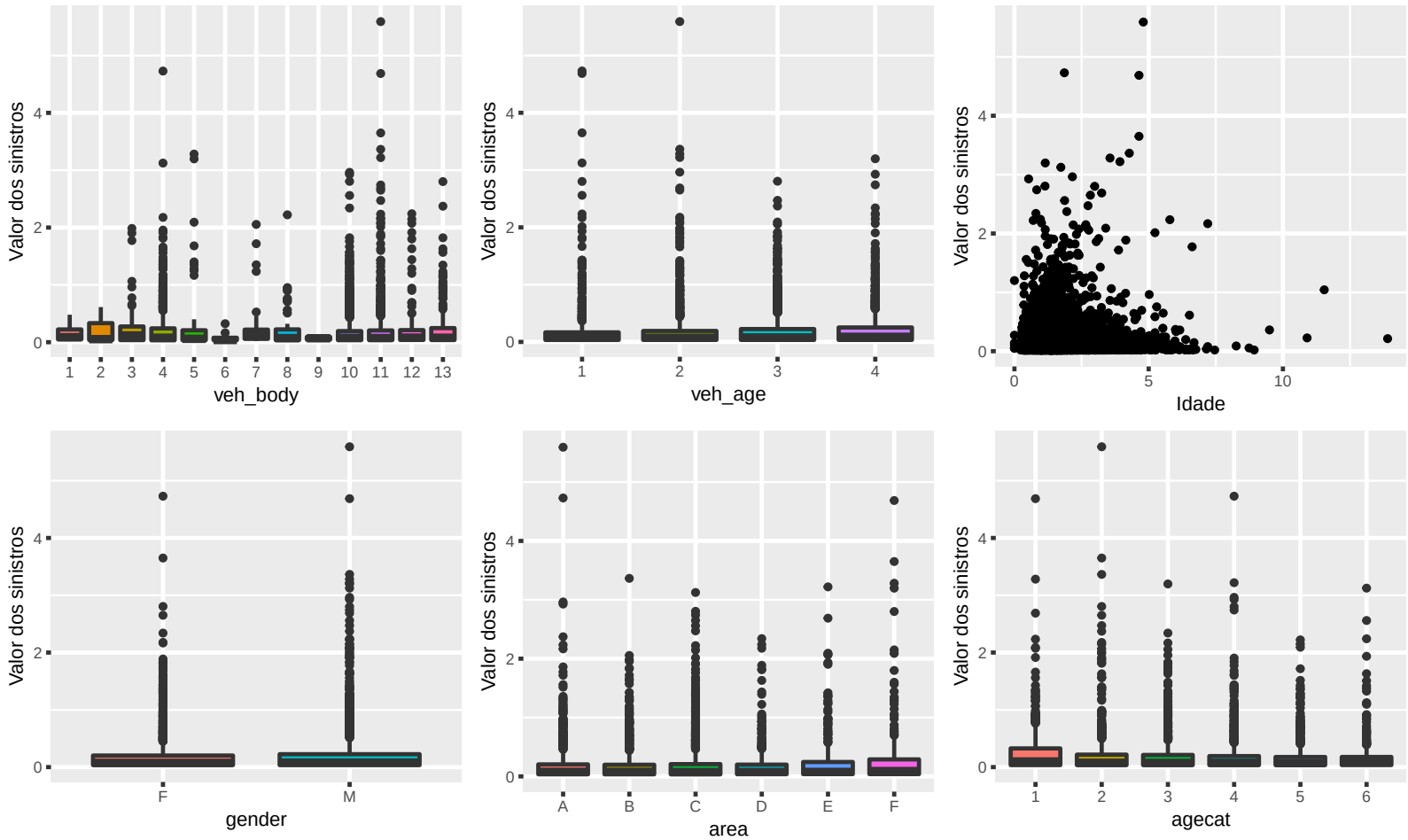

Figura 13: Distribuição dos valores de sinistros segundo as covariáveis

Exemplo 6 - Sinistros em apólices de seguros de automóveis

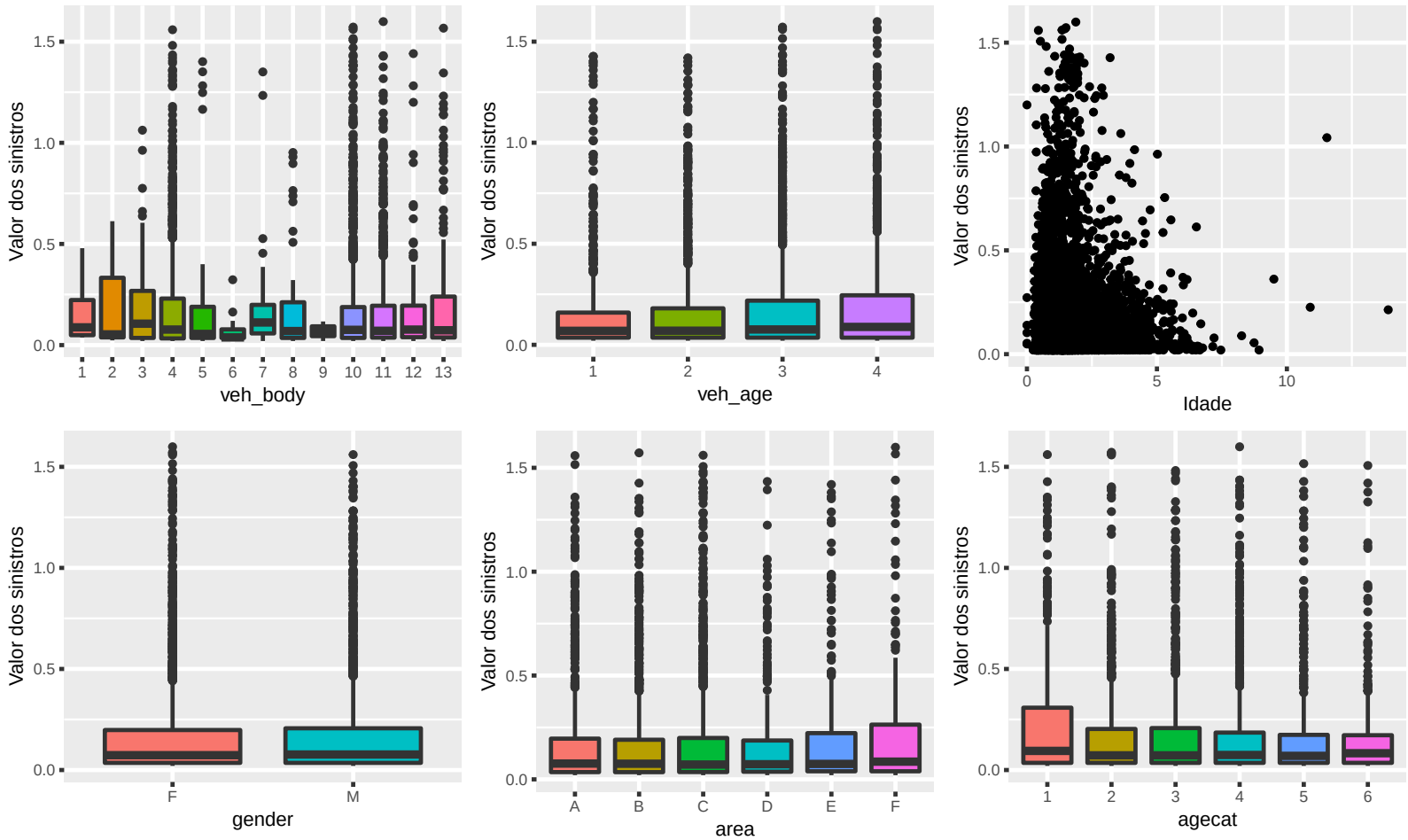

Figura 14: Distribuição dos valores de sinistros segundo as covariáveis (desconsiderando sinistros superiores a 15.000 dólares.

Exemplo 6 - Sinistros em apólices de seguros de automóveis

- **Objetivos da análise:**
- Identificar fatores associados a maiores valores de sinistros;
- Estabelecer um modelo para precificação de apólices.

Mãos a obra!