Modelos Lin. Generalizados - Lista 3

Ângelo Majeau RA: 727843
Clézio Lopes RA: 727849
Mayara Formenton RA: 632023

Exercício 1

Inicialmente, será verificada se há uma possível relação entre a variável resposta (Faturamento anual) e a preditora (Gasto com propaganda).

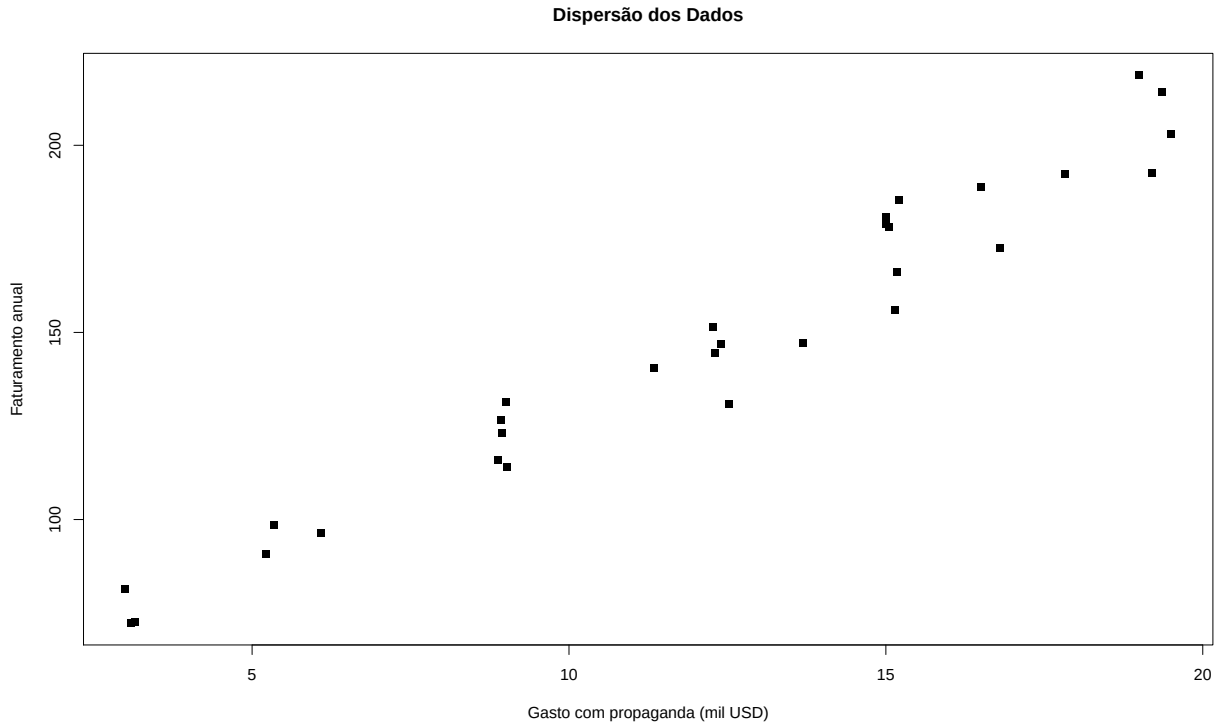

Figura 1: Faturamento vs Gasto

Nota-se pela Figura 1 que existe uma relação linear positiva entre a variável faturamento e gasto com propaganda, ou seja, quanto maior o gasto com propaganda maior o faturamento médio do restaurante.

Modelo 1

O modelo 1 é dado por uma regressão linear simples, ou um modelo linear generalizado com distribuição dos erros normal e função de ligação dada pela função identidade.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

em que Y_i corresponde ao valor médio de faturamento dos restaurantes e X_i é o gasto com propaganda.

Tabela 1: Estimativas do modelo 1

Parâmetro	Estimativa	DP	Valor t	p-valor
Intercepto	49.4434	4.2889	11.53	3.81e-12
Gasto	8.0484	0.3265	24.65	< 2e-16

Para este modelo temos o diagnóstico apresentado na Figura 2.

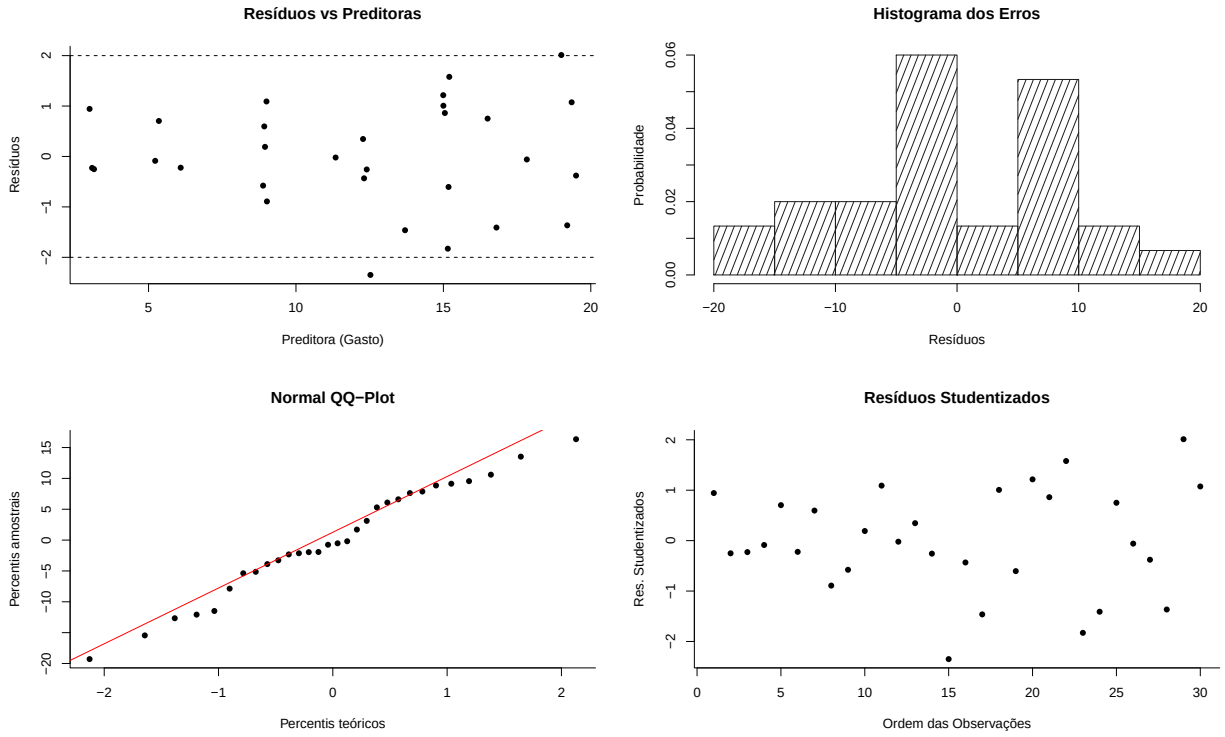

Figura 2: Diagnóstico modelo 1

Pelos gráficos apresentados na Figura 2 tem-se que os erros estão distribuídos de maneira independente, mas pelo gráfico de resíduos por ordem das preditoras observa-se que a variância não é constante, já que até a observação 15 temos uma variância baixa, enquanto que a partir desse ponto a variância aumenta consideravelmente. Sendo assim, serão ajustados outros modelos.

Outros Modelos

Além do modelo citado anteriormente (regressão linear simples), foram ajustados outros modelos com erros Gama e erros Gaussiana Inversa, ambos com função de ligação *log* e identidade.

Tabela 2: AIC's dos modelos ajustados

Modelo	AIC
Modelo (Normal,identidade)	220.89
Modelo (Gamma,log)	225.35
Modelo (Gamma,identidade)	215.96
Modelo (Gau. Inversa,log)	227.84
Modelo (Gau. Inversa,identidade)	216.88

Pelos *AIC* apresentados na Tabela 2 temos que um bom modelo para o conjunto de dados terá função de ligação *log*, pois com ela temos que os *AIC* dos modelos são mais baixos.
Vamos ao diagnóstico do modelo Gama com ligação identidade pois é a que apresenta o menor *AIC*.

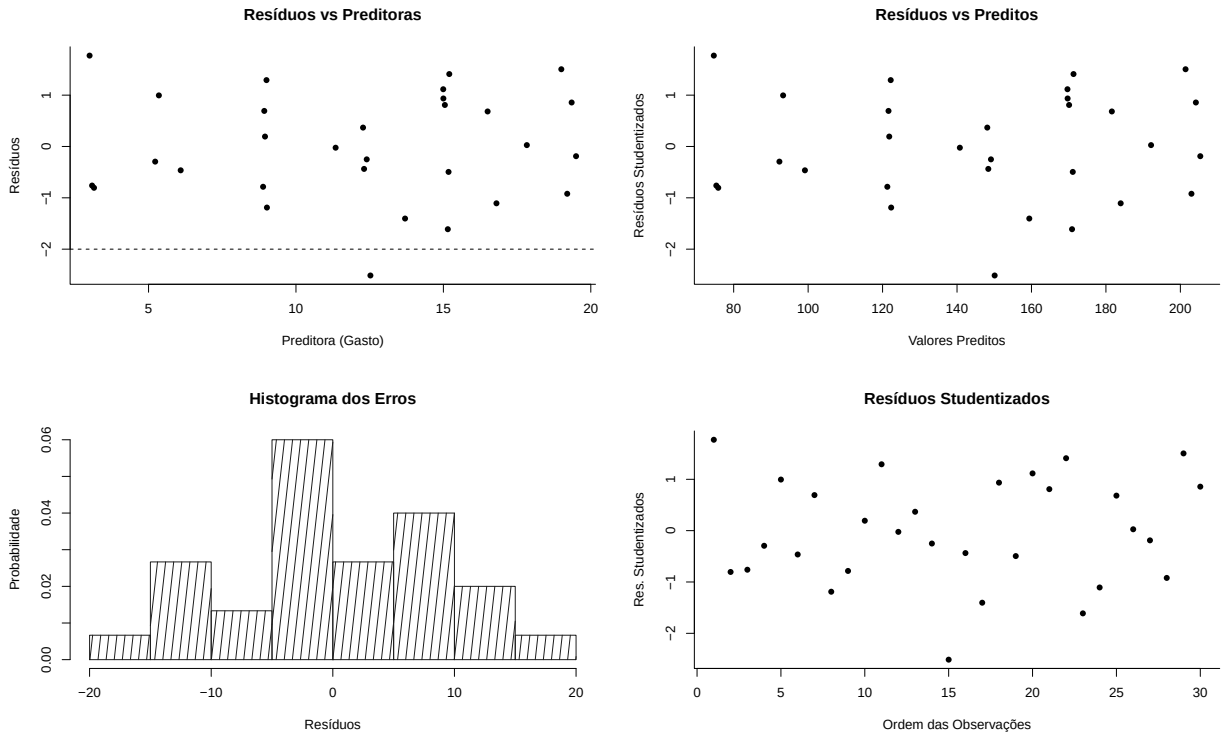

Figura 3: Diagnostico do modelo Gama-identidade

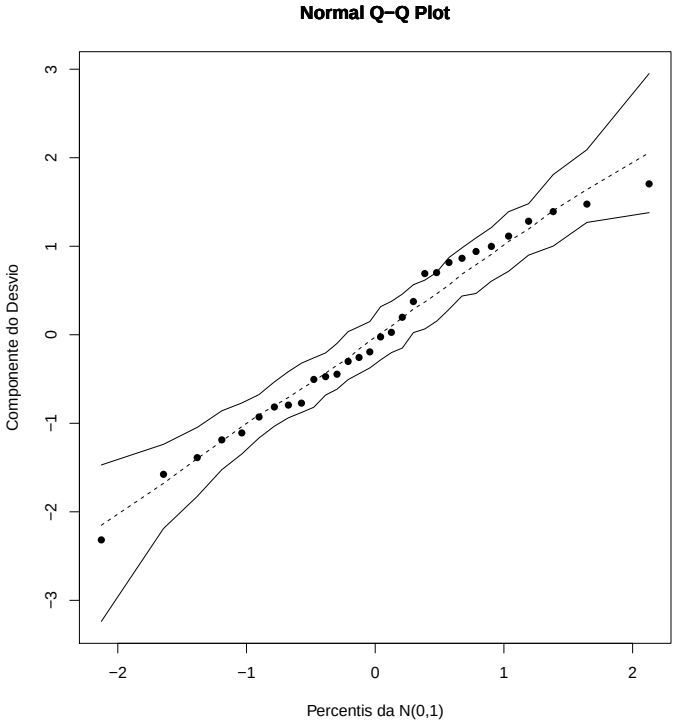

Figura 4: Envelope simulado do modelo

Observa-se pelos gráficos na Figura 3 que as suposições para este modelo estão atendidas. Embora na Figura 4 possua um ponto fora do envelope simulado, o mesmo ainda é muito próximo das bandas de confiança, então não podemos dizer que este modelo é inadequado. Desta forma o modelo escolhido é dado por erros com distribuição Gama com função de ligação *log*.

Interpretação dos Parâmetros

As estimativas para o modelo escolhido (Gama com ligação identidade) é dada por:

Tabela 3: Estimativas do modelo 3

Parâmetro	Estimativa	DP	Valor t	p-valor
Intercepto	50.9029	2.6855	18.95	<2e-16
Gasto	7.9205	0.2644	29.95	<2e-16

Estima-se que o faturamento médio dos restaurante é igual a 50.9 mil dólares (USD) quando não há gasto com propaganda, e ainda a medida que aumenta uma unidade de gasto com propaganda (mil USD), há um aumento de 7.92 mil dólares no faturamento médio dos restaurantes.

Exercício 3

Tabela 4: Descrição das variáveis

Variável	Descrição	Tipo
Y	consumo de combustível (milhas por galão)	contínua
X_2	número de cilindros	discreta
X_3	cilindradas	contínua
X_5	peso	contínua
X_6	aceleração	contínua
X_7	ano do modelo	discreta
X_8	origem	categórica

Diante das variáveis obtidas, o objetivo é prever o consumo de combustível de veículos em milhas por galão, em termos das variáveis preditoras.

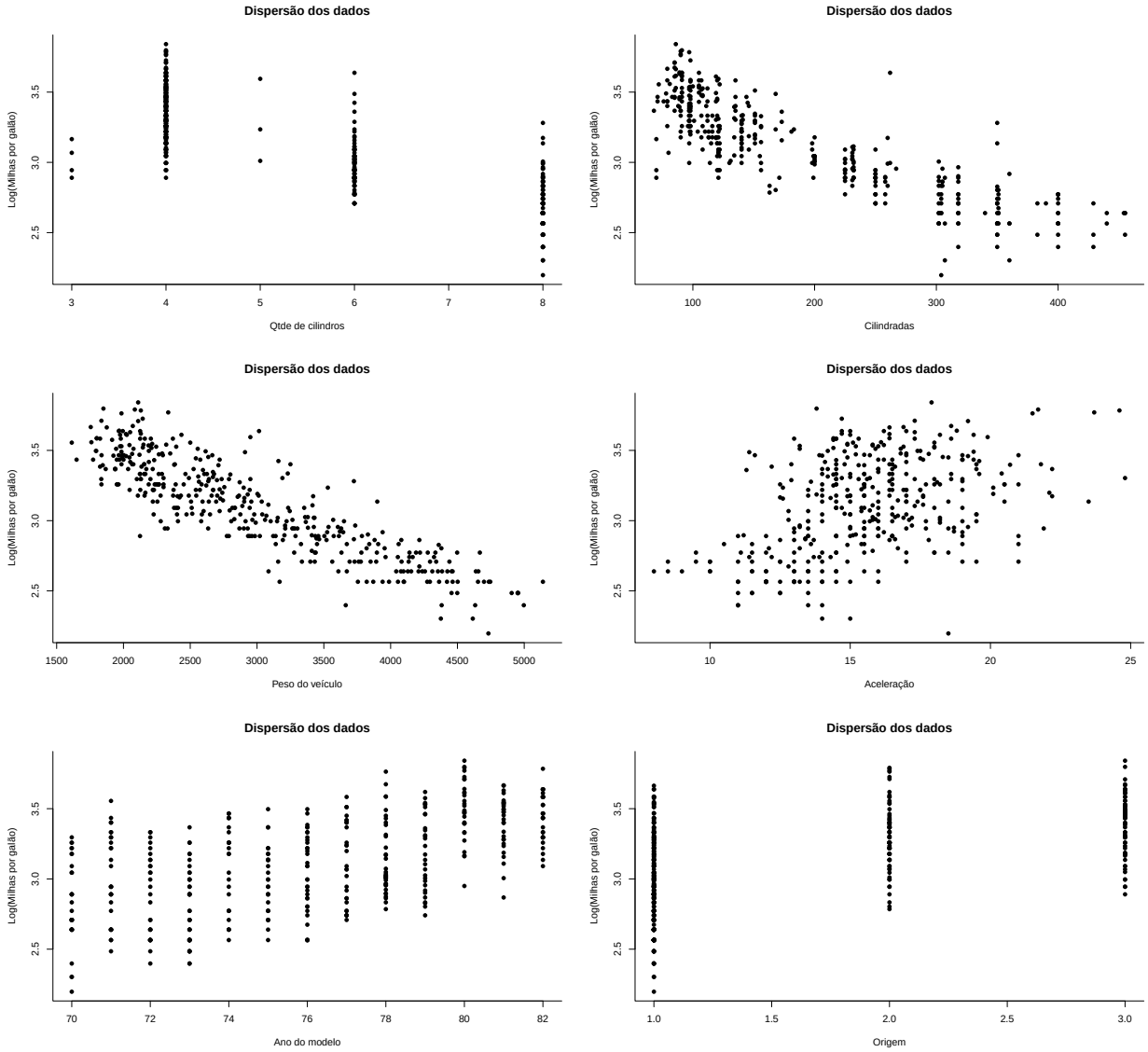

Figura 5: Gráfico de dispersões por variável preditora

a) Por meio da Figura 5, verifica-se que conforme aumenta a cilindrada, o *log(Y)* diminui, isto é, há uma relação linear negativa. O mesmo comportamento acontece para a variável peso, ou seja, quanto mais pesado o veículo, menor será o *log* do consumo de combustível. Já para a variável aceleração, verifica-se que quanto mais veloz é o veículo, maior será o *log(Y)*. Quanto a variável origem, a medida que ela passa do fator 1 para o 2 ou 3, o *log(Y)* também aumenta, o que indica que há uma relação entre o *log(Y)* e as variáveis preditoras.

   Para a variável quantidade de cilindro podemos observar que carros com 8 cilindros tem o *log(Y)* menor em relação aos demais níveis, e para a variável ano de modelo do veículo observa-se que quanto maior o ano maior o *log(Y)*, dando indícios de que quanto mais novo é o veículo maior é o *log(Y)*. Sendo assim, é recomendado o modelo de regressão com função de ligação logarítmica.

b) Como a variável X_8 é qualitativa, serão criadas variáveis *dummyes* com duas classes para representar os níveis da origem. Sendo assim, tem-se:

$$X_{8.2} = \begin{cases} 1, & \text{se é do segundo nível} \\ 0, & \text{caso contrario} \end{cases}$$

$$X_{8.3} = \begin{cases} 1, & \text{se é do terceiro nível} \\ 0, & \text{caso contrario} \end{cases}$$

Portanto, o primeiro nível foi utilizado como referência.

O modelo inicial será constituído por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_2 + \hat{\beta}_2 X_3 + \hat{\beta}_3 X_5 + \hat{\beta}_4 X_6 + \hat{\beta}_5 X_7 + \hat{\beta}_6 X_{8.2} + \hat{\beta}_7 X_{8.3}$$

A fim de saber se as variáveis são significativas para o modelo, será testado um modelo através do método *stepwise*, em que será ajustado uma sequência de modelos de regressão, em cada passo é adicionada ou excluída uma variável preditora X do modelo.

Tabela 5: Passo a passo das entradas do método *stepwise*

Passo	**Variável**	**AIC**
1	X_5	2170.2
2	X_7	1929.6
3	X_8	1912.7
4	X_3	1914.4
5	X_6	1912.6

Através de tal método de seleção, obteve-se o seguinte conjunto de variáveis preditoras, respectivamente: X_5, X_7, X_8, X_3, X_6.

c) O modelo final obtido no item anterior é dado por:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_3 + \hat{\beta}_2 X_5 + \hat{\beta}_3 X_6 + \hat{\beta}_4 X_7 + \hat{\beta}_5 X_{8.2} + \hat{\beta}_6 X_{8.3}$$

Os valores estimados pelo modelo podem ser visto na tabela a seguir:

Tabela 6: Estimativas do modelo

Variável	**Par. Estimado**	**S. Error**	**t value**	**p-valor**
Intercepto	1.420	1.463e-01	9.711	< 2e-16
X_3	2.769e-04	1.982e-04	1.397	0.163187
X_5	-3.070e-04	2.073e-05	-14.808	< 2e-16
X_6	5.314e-03	2.735e-03	1.943	0.052706
X_7	3.204e-02	1.782e-03	17.978	< 2e-16
$X_{8.2}$	8.476e-02	1.964e-02	4.316	2.02e-05
$X_{8.3}$	6.843e-02	1.899e-02	3.603	0.000355

O modelo final teve um AIC de 1912.6. Para fazer as interpretações dos parâmetros é necessário aplicar a exponencial nos valores estimados, pois estamos trabalhando com função ligação logarítmica.

- $exp\{\hat{\beta}_1\} = exp\{0.0002769\} = 1.000277$: Estima-se que a média do consumo de combustível por milhas por galão sofra um acréscimo de 0.0277% quando acrescenta-se uma unidade na variável número de cilindros e mantém-se as demais variáveis preditoras constantes.
- $exp\{\hat{\beta}_2\} = exp\{-0.0003070\} = 0.999693$: Estima-se que a média do consumo de combustível por milhas por galão sofra um decréscimo de 0.0306% quando acrescenta-se uma unidade na variável peso e mantém-se as demais variáveis preditoras constantes.
- $exp\{\hat{\beta}_3\} = exp\{0.005314\} = 1.005328$: Estima-se que a média do consumo de combustível por milhas por galão sofra um acréscimo de 0.533% quando acrescenta-se uma unidade na variável aceleração e mantém-se as demais variáveis preditoras constantes.
- $exp\{\hat{\beta}_4\} = exp\{0.03204\} = 1.032559$: Estima-se que a média do consumo de combustível por milhas por galão sofra um acréscimo de 3.25% quando acrescenta-se uma unidade na variável ano do modelo e mantém-se as demais variáveis preditoras constantes.
- $exp\{\hat{\beta}_5\} = exp\{0.08476\} = 1.088456$: Estima-se que a média do consumo de combustível por milhas por galão sofra um acréscimo de 8.84% é passado do nível 1 (referência) da variável origem para o nível 2, mantendo as demais variáveis preditoras constantes.
- $exp\{\hat{\beta}_6\} = exp\{0.006843\} = 1.070826$: Estima-se que a média do consumo de combustível por milhas por galão sofra um acréscimo de 7.08% é passado do nível 1 (referência) da variável origem para o nível 3, mantendo as demais variáveis preditoras constantes.

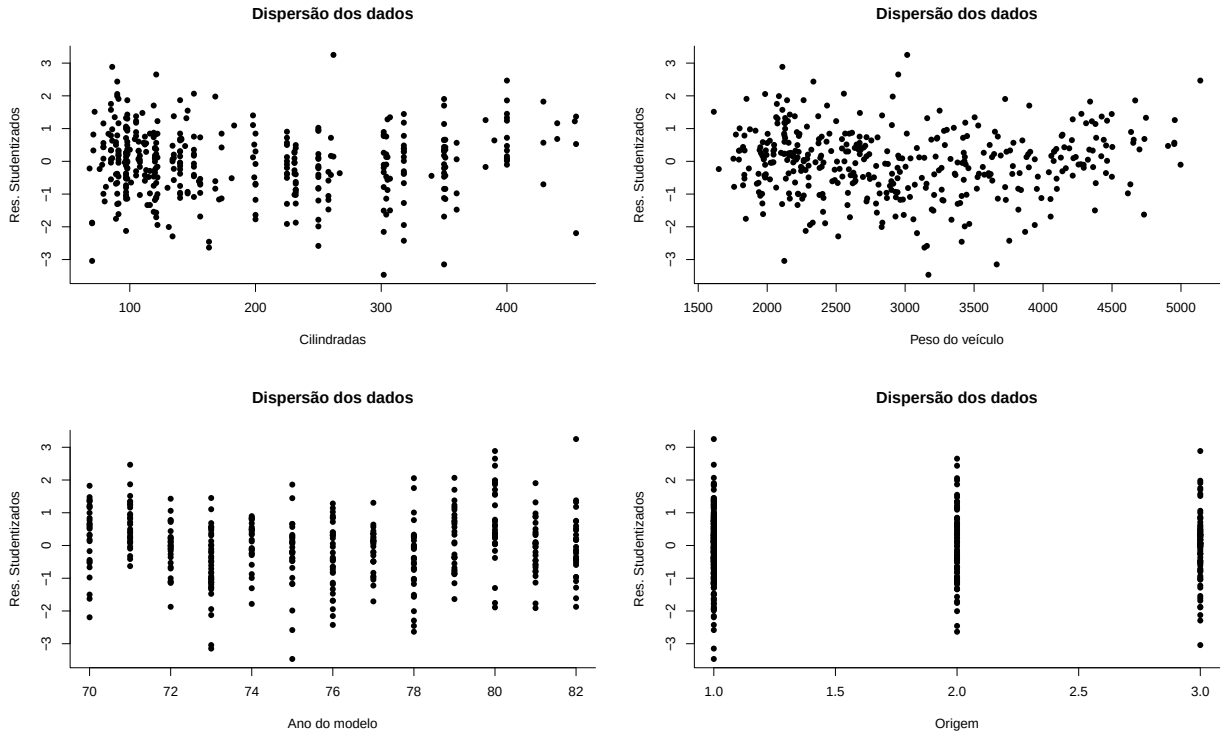

Figura 6: Resíduos por variável preditora

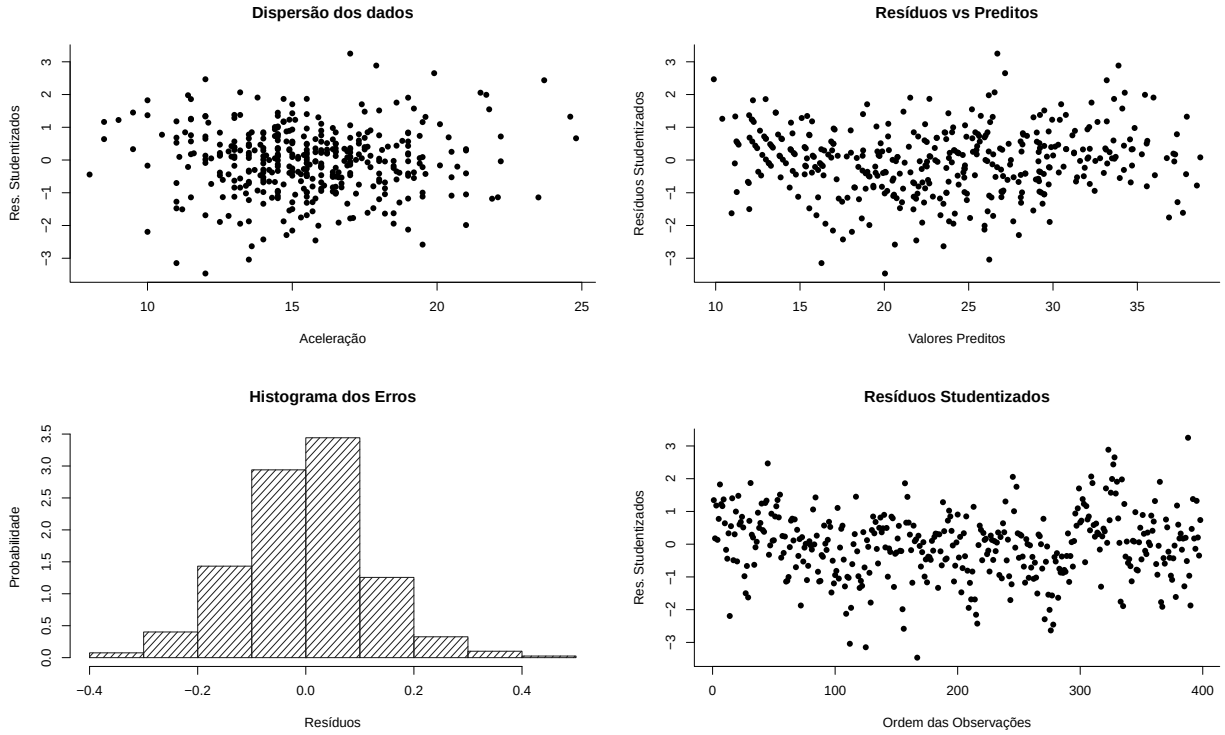

Figura 7: Resíduos por variável preditora

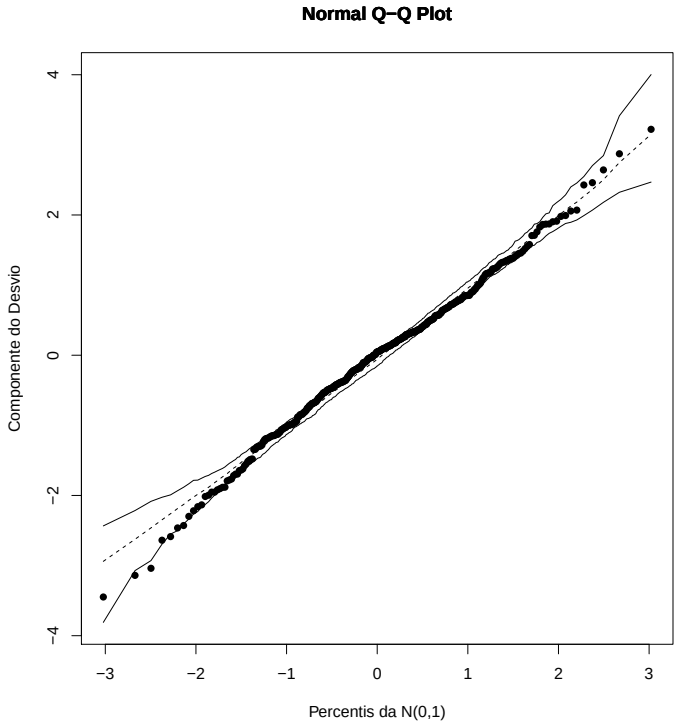

Figura 8: Gráfico de dispersões por variável preditora

d) Tem-se que a primeira suposição de que $Y_i|X_i$ é independente de $Y_j|X_j$ é atendida, pois através do gráfico de resíduos x ordem das observações, nota-se que ocorre um comportamento aleatório.
É possível verificar no gráfico de resíduos x valores preditos um comportamento aleatório, o que indica que ϕ não varia em função das variáveis preditoras.
Para verificar se a suposição de que $Y_i|X_i$ tem distribuição pertencente à Família Exponencial

Linear e de que a distribuição para Y_i é correta, é necessário que os resíduos estejam contidos na banda de confiança do envelope simulado. Observa-se então por meio do gráfico de probabilidade normal que a suposição é atendida, uma vez que não contém pontos muito distantes das bandas de confiança.
Nota-se também que a função de ligação escolhida é correta, pois verifica-se um comportamento aleatório dos resíduos tanto para o gráfico de resíduos x valores preditos quanto para os gráficos de resíduos x cada uma das preditoras. Tal comportamento aleatório no gráfico de resíduos x cada uma das preditoras também indica que todas as variáveis foram inseridas corretamente no modelo. Diante disso, não foi constatado problemas com a análise de diagnóstico, pois todas foram atendidas.

e) Para a análise de pontos alavancas, discrepantes e influentes temos.

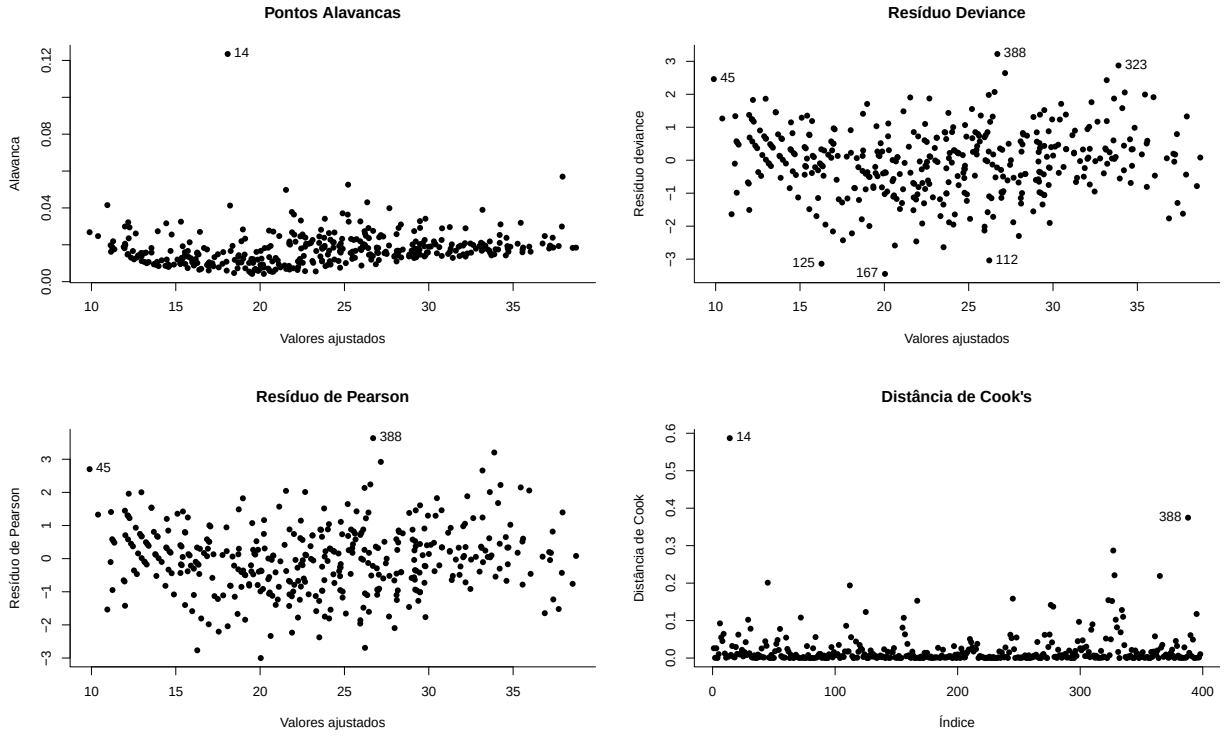

Figura 9: Gráfico de pontos Alavancas, Discrepantes e Influentes

Por meio do gráfico de alavanca x valores ajustados, percebe-se que o ponto 14 é alavanca, uma vez que este está muito distante dos demais. Além disso, o gráfico da Distância de Cook's revela que o mesmo ponto também é influente, levando em conta que a distância dele é grande comparada com os demais. Nota-se nos gráficos de resíduo deviance e resíduo de Pearson que não há pontos discrepantes.
Portanto, será retirado o ponto 14 a fim de verificar se as estimativas dos coeficientes do modelo são alteradas.

Tabela 7: Estimação dos parâmetros com e sem a observação 14

Variável	Par. Estimado (com 14)	Par. Estimado (sem 14)
Intercept	1.420	1.426
X_3	2.769e-04	4.163e-04
X_5	-3.070e-04	-3.215e-04
X_6	5.314e-03	5.702e-03
X_7	3.204e-02	3.208e-02
$X_{8.2}$	8.476e-02	8.843e-02
$X_{8.3}$	6.843e-02	7.019e-02

Observa-se através da Tabela 7 que não há grandes alterações nas estimativas dos parâmetros quando retira-se a observação 14, exceto para o parâmetro associado a variável X_3. Entretanto, como a mudança significativa no parâmetro acontece apenas para uma variável, não há necessidade de remover tal observação.

f) Para obter o psuedo R^2 do modelo, utiliza-se a seguinte fórmula:

$$\begin{aligned}
\frac{R^2_{CSM}}{max(R^2_{CSM})} &= \frac{1 - exp\{-\frac{2(l(\hat{\gamma}) - l(0))}{n}\}}{1 - exp\{\frac{2l(0)}{n}\}} \\
&= \frac{1 - exp\{-\frac{2(-947.3-(-1367))}{398}\}}{1 - exp\{\frac{2(-1367)}{398}\}} \\
&= \frac{1 - 0.1213538}{1 - 0.001039156} \\
&= 0.8795602
\end{aligned}$$

Assim, entende-se que o modelo final explica aproximadamente 87.95% da variabilidade do consumo de combustível por milhas por galão.

Exercício 4

a) Para testar se a média da variável resposta varia em função dos níveis de X_8, realizamos o teste da razão de máxima verossimilhança entre os modelo com e sem a variável preditora em questão.

$$\begin{cases} H_0 : (\beta_5, \beta_6) = (0, 0) \text{ (A média de mpg não varia em função da variável qualitativa } X_8) \\ H_1 : (\beta_5, \beta_6) \neq (0, 0) \text{ (A média de mpg varia em função da variável qualitativa } X_8) \end{cases}$$

Tabela 8: Teste da razão de máxima verossimilhança

Modelo	LogLik	Df	Chisq	p-valor
1	-948.29	8		
2	-959.18	6	21.783	1.862e-05

Ao nível de significância usual de 5% rejeita-se H_0, isto é, há evidências que a média do consumo de combustível varia em função da variável qualitativa X_8.

b)

$$\begin{cases} H_0 : \beta_5 = 0 \text{ (A média de mpg não varia em função dos níveis 1 e 2 da variável qualitativa } X_8) \\ H_1 : \beta_5 \neq 0 \text{ (A média de mpg varia em função dos níveis 1 e 2 da variável qualitativa } X_8) \end{cases}$$

Tabela 9: Teste do parâmetro $\hat{\beta}_5$

Teste	**Chisq**	**p-valor**
$\hat{\beta}_5 = 0$	18.624	1.592e-05

Pode-se analisar que ao nível de significância de 5% há evidências que a média de milhas percorridas por galão dos veículos em estudo varia em função dos níveis 1 e 2 da variável origem do carro.

$$\begin{cases} H_0 : \beta_6 = 0 \text{ (A média de mpg não varia em função dos níveis 1 e 3 da variável qualitativa } X_8) \\ H_1 : \beta_6 \neq 0 \text{ (A média de mpg varia em função dos níveis 1 e 3 da variável qualitativa } X_8) \end{cases}$$

Tabela 10: Teste do parâmetro $\hat{\beta}_6$

Teste	**Chisq**	**p-valor**
$\hat{\beta}_6 = 0$	12.981	0.0003147

Analisando o p-valor e utilizando o nível de confiança de 95% temos evidências de que o a média de milhas percorridas por galão varia em função dos níveis 1 e 3 da variável preditora origem, sendo assim devem continuar no ajuste do modelo.

$$\begin{cases} H_0 : (\beta_5 - \beta_6) = 0 \text{ (A média de mpg não varia em função dos níveis 2 e 3 da variável qualitativa } X_8) \\ H_1 : (\beta_5 - \beta_6) \neq 0 \text{ (A média de mpg varia em função dos níveis 2 e 3 da variável qualitativa } X_8) \end{cases}$$

Tabela 11: Teste da diferença $(\hat{\beta}_5 - \hat{\beta}_6)$

Teste	**Chisq**	**p-valor**
$\hat{\beta}_5 - \hat{\beta}_6 = 0$	0.6731	0.412

Pela Tabela 11 temos evidências de que a diferença entre os parâmetros $\hat{\beta}_5$ e $\hat{\beta}_6$ não é igual a zero, ou seja, a média de milhas por galão não varia em função dos níveis 2 e 3 da variável origem, com *p-valor* $= 0.412$ não temos evidências para rejeitar H_0.

c) Para obter esse intervalo de confiança para a razão entre a média da variável resposta entre os níveis 2 e 3, foi escolhido um alfa de 5%. O resultado é obtido a partir da fórmula:

$$\begin{aligned}
\text{IC}(a^t\beta; 1-\alpha) &= a^t\hat{\beta} \pm z_{1-\frac{\alpha}{2}}\sqrt{a^t(K_(\beta\beta))^{-1}a} \\
&= \hat{\beta}_5 - \hat{\beta}_6 \pm 1.96(0.0003961194) \\
&= \hat{\beta}_5 - \hat{\beta}_6 \pm 1.96(0.0003961194) \\
&= (-0.02267939; 0.05533939)
\end{aligned}$$

Em termos de interpretação, é necessário aplicar a exponencial nos limites encontrados do intervalo, chegando no seguinte resultado: (0.978 ; 1.057). Portanto, nota-se que o valor 1 está no intervalo. Sendo assim, por se tratar de uma razão podem-se concluir que a média da variável resposta não muda quando a variável origem passa do nível 2 para o nível 3. Observa-se que esse resultado era o esperado depois do teste de hipótese realizado no item b).

d) Com os seguintes valores: x2 = 4, x3 = 150, x4 = 100, x5 = 2300, x6 = 17, x7 = 80, x8 = 1, mas descartando as informações referentes a X_2 e X_4 pois não estão presentes no modelo, temos a seguinte estimativa:

$$\begin{aligned}
\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_3 + \hat{\beta}_2 X_5 + \hat{\beta}_3 X_6 + \hat{\beta}_4 X_7 + \hat{\beta}_5 X_{8.2} + \hat{\beta}_6 X_{8.3} \\
&= 1.420 + 0.000277 X_3 - 0.00031 X_5 + 0.005314 X_6 + 0.03204 X_7 + 0.08476 X_{8.2} + 0.06843 X_{8.3} \\
&= 1.420 + 0.000277(150) - 0.00031(2300) + 0.005314(17) + 0.03204(80) + 0.08476(0) + 0.06843(0) \\
&= 1.420 + 0.000277(150) - 0.00031(2300) + 0.005314(17) + 0.03204(80) \\
&= 1.420 + 0.041535 - 0.7061 + 0.090338 + 2.5632 \\
\hat{Y}_i &= 3.409072
\end{aligned}$$

Temos que o intervalo para $\mathbb{E}(\hat{y}_i)$ é dado por:

$$\begin{aligned}
\text{IC}(\mathbb{E}(\hat{y}_i; 1-\alpha) &= x_0^t\hat{\beta} \pm z_{1-\frac{\alpha}{2}}\sqrt{x_0^t(K_(\beta\beta))^{-1}x_0} \\
&= 3.409072 \pm 1.96(0.01375822) \\
&= 3.409072 \pm 1.96(0.01375822) \\
&= (3.395313; 3.42283)
\end{aligned}$$

Para realizar a interpretação, é necessário aplicar a exponencial nos limites encontrados do intervalo, chegando no seguinte resultado: (29.824 ; 30.656). Portanto, podemos concluir que a verdadeira média da variável resposta quando fixamos os valores das variáveis preditoras conforme mencionamos acima está contida no intervalo (29.824 ; 30.656) com 95% de confiança.

e) Para realizar a categorização da Variável X_7 a melhor maneira segundo o grupo é:

$$X_7 = \begin{cases} 1, & \text{se é dos anos (70, 71, 72, 73 e 74)} \\ 2, & \text{se é dos anos (75, 76, 77, 78 e 79)} \\ 3, & \text{se é dos anos (80, 81 e 82} \end{cases}$$

Desta maneira, evitando criar muitas varíaveis *dummys* e como um critério razoável sendo agrupar de cinco em cinco anos onde uma diferença de mais cinco anos possa ser significativa enquanto o contrário não, além disso, foram criadas mais duas variáveis *dummys* para entrar no modelo e representar as categorias da variável X_7, onde o nível de referência é o 1.

Após a categorização da variável X_7 observamos que algumas variáveis deixaram de ser significativas, sendo assim, foi realizado um novo conjunto de seleção de variáveis e o modelo é dado por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_5 + \hat{\beta}_2 X_{7.2} + \hat{\beta}_3 X_{7.3} + \hat{\beta}_4 X_{8.2} + \hat{\beta}_5 X_{8.3}$$

Com estimativas dadas por:

Tabela 12: Estimativas do modelo

Variável	Par. Estimado	S. Error	t value	p-valor
Intercepto	3.821	3.388e-02	112.800	<2e-16
X_5	-2.880e-04	9.186e-06	-31.356	<2e-16
$X_{7.2}$	1.324e-01	1.365e-02	9.699	<2e-16
$X_{7.2}$	3.140e-01	1.698e-02	18.489	<2e-16
$X_{8.2}$	6.579e-02	1.821e-02	3.613	0.000343
$X_{8.3}$	4.519e-02	1.850e-02	2.443	0.015011

As interpretações seguem o mesmo padrão já comentado no Exercício 3.

f) Adicionando interação no modelo pela qualitativa com as demais preditoras temos o seguinte ajuste:

Tabela 13: Estimativas do modelo com interação

Variável	Par. Estimado	S. Error	t value	p-valor
Intercepto	3.756	4.974e-02	75.517	< 2e-16
X_5	-2.704	1.326e-05	-20.388	< 2e-16
$X_{7.2}$	3.164e-01	7.122e-02	4.442	1.17e-05
$X_{7.3}$	3.431e-01	1.013e-01	3.387	0.00078
$X_{8.2}$	8.432e-02	3.111e-02	2.710	0.00703
$X_{8.3}$	8.208e-02	3.419e-02	2.401	0.01683
$X_5 : X_{7.2}$	-4.985e-05	1.995e-05	-2.499	0.01286
$X_5 : X_{7.3}$	-1.554e-05	3.472e-05	-0.447	0.65481
$X_{7.2} : X_{8.2}$	-9.361e-02	4.149e-02	-2.256	0.02462
$X_{7.3} : X_{8.2}$	9.042e-02	4.804e-02	1.882	0.06057
$X_{7.2} : X_{8.3}$	-9.225e-02	4.627e-02	-1.994	0.04688
$X_{7.2} : X_{8.3}$	-6.634e-03	4.607e-02	-0.144	0.88559

Pela Tabela 13 temos que ao nível de significância de 5% temos que somente as interações entre X_5 e $X_{7.2}$, X_5 e $X_{7.3}$ e as $X_{7.2}$ e $X_{8.2}$, $X_{7.2}$ e $X_{8.3}$ devem ser mantidas no modelo, as demais interações podem ser removidas.

Vale ressaltar que o modelo com interações e X_7 categorizada possui um *AIC* igual a 1910.6 que é inferior ao do modelo quando X_7 era não categórica, desta forma um bom modelo para prever a quantidade média de milhas percorridas por galão deve possuir X_7 categorizadas e com algumas interações.