Resolução da segunda lista de exercícios do curso MAE0328 disponível no Texto de Regressão (Paula, 2023).

10. Considere o seguinte modelo de regressão linear múltipla para um determinado conjunto de p variáveis $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$, em que $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, $i = 1, \ldots, n$. Sabemos que a quantidade avaliada segundo o critério de Akaike é dada por

$$-2\text{log}\{L(\hat{\boldsymbol{\beta}}, \hat{\sigma};\ \mathbf{y})\} + 2p,$$

em que $\hat{\boldsymbol{\beta}}$ e $\hat{\sigma}$ são as estimativas de máxima verossimilhança de $\boldsymbol{\beta}$ e σ, e $L(\cdot)$ é a função de verossimilhança do modelo. Note que $\hat{\sigma}^2 = \text{SQRes}/n$, logo temos que

$$\begin{aligned}
-2\text{log}\{L(\hat{\boldsymbol{\beta}}, \hat{\sigma};\ \mathbf{y})\} &= -2\text{log}\left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \text{exp}\left\{-\frac{1}{2\hat{\sigma}^2}(y_i - \hat{y}_i)^2\right\}\right] \\
&= n\text{log}(2\pi\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= n\text{log}\left(\frac{\text{SQRes}}{n}\right) + n\text{log}(2\pi) + n.
\end{aligned}$$

Como $n\text{log}(2\pi) + n$ é constante para qualquer outro conjunto de variáveis, um modelo encaixado escolhido segundo o critério de Akaike tem o menor valor de $n\text{log}\,(\text{SQRes}/n) + 2p$ entre os modelos avaliados.

11. Considere o conjunto de dados apresentado na Seção 6.8 de Lawless (1982) em que são apresentados os resultados de experimentos para avaliar a resistência de um determinado tipo de vidro e são descritas as seguintes variáveis: `Time`: Tempo de resistência em horas; `Volt`: 1 (200kV), 2 (250kV), 3 (300kV) ou 4 (350kV) para os níveis de voltagem do experimento e `Temp`: 1 (170ºC) ou 2 (180ºC) para os níveis de temperatura do experimento. A Figura 1 mostra os perfis médios com bandas de erro padrão da resistência segundo a voltagem para os dois níveis de temperatura, verificamos descritivamente indícios de interação entre temperatura e voltagem devido ao fato do efeito da voltagem na resistência do vidro não ser o mesmo dependendo do nível da temperatura.

Seja $\texttt{Time}_{ijk}$ o tempo de resistência da k-ésima amostra de vidro submetida à i-ésima voltagem e à j-ésima temperatura, supondo inicialmente o seguinte modelo

$$\texttt{Time}_{ijk} = \alpha + \beta_i + \gamma_j + \epsilon_{ijk}, \quad \epsilon_{ijk} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2),$$

em que β_i denota o efeito da i-ésima voltagem e γ_j o efeito da j-ésima temperatura, ambos em relação à casela de referência, sendo assumido

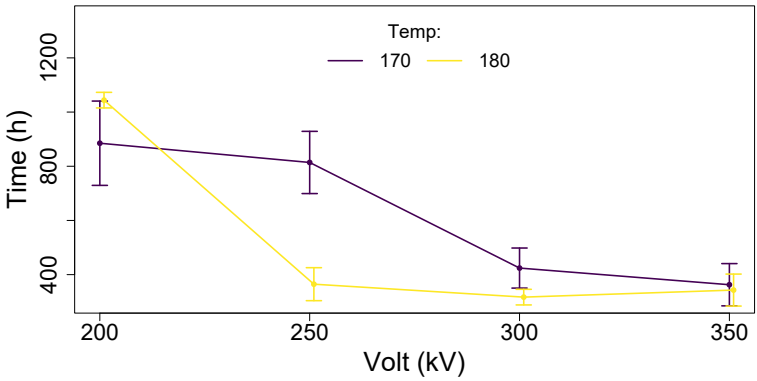

Figura 1 Perfis médios com bandas de erro padrão da resistência segundo a voltagem para os dois níveis de temperatura.

$\beta_1 = 0$, $\gamma_1 = 0$ para $i = 1, \ldots, 4$ e $j = 1, 2$. A Tabela 1 apresenta a tabela ANOVA do ajuste do modelo com interação, notamos que o efeito de interação entre temperatura e voltagem é significativo, avaliado com nível de 10%, logo foi mantido no modelo.

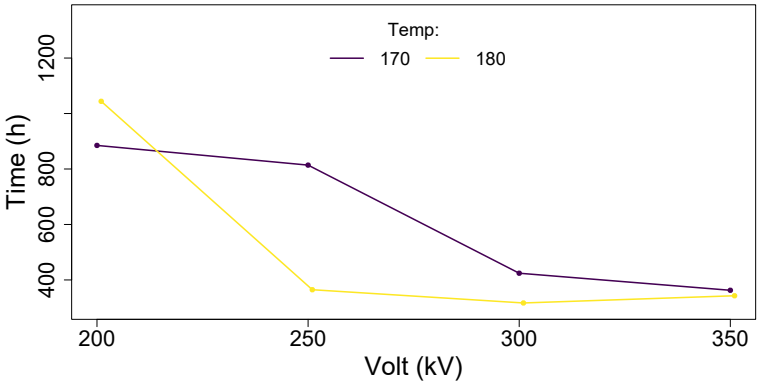

Figura 2 Perfis ajustados do modelo para resistência segundo a voltagem e a temperatura.

A Figura 3 mostra os gráficos da análise de resíduos e de sensibilidade do modelo ajustado, verificamos que as suposições de homocedasticidade da variância, independência e normalidade das observações parecem satisfeitas. O gráfico da distância de Cook destaca a observação #1 com tempo de resistência de 439 horas, que é muito menor que as demais para voltagem de 200 kV. Como análise confirmatória, removendo tal observação o valor-P do teste da interação muda para 0.0109. Portanto, como

Tabela 1 Tabela ANOVA do modelo ajustado com interação para o tempo de resistência do vidro segundo a voltagem e a temperatura.

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	valor-F	valor-P
`Volt`	3	1943181	647727	22.41	< 0.0001
`Temp`	1	87049	87049	3.01	0.0954
`Volt:Temp`	3	390950	130317	4.51	0.0120
Total	24	693486	28895		

não há mudança de decisão e as suposições estão satisfeitas, a análise de diagnóstico validou a decisão tomada a respeito do efeito de interação.

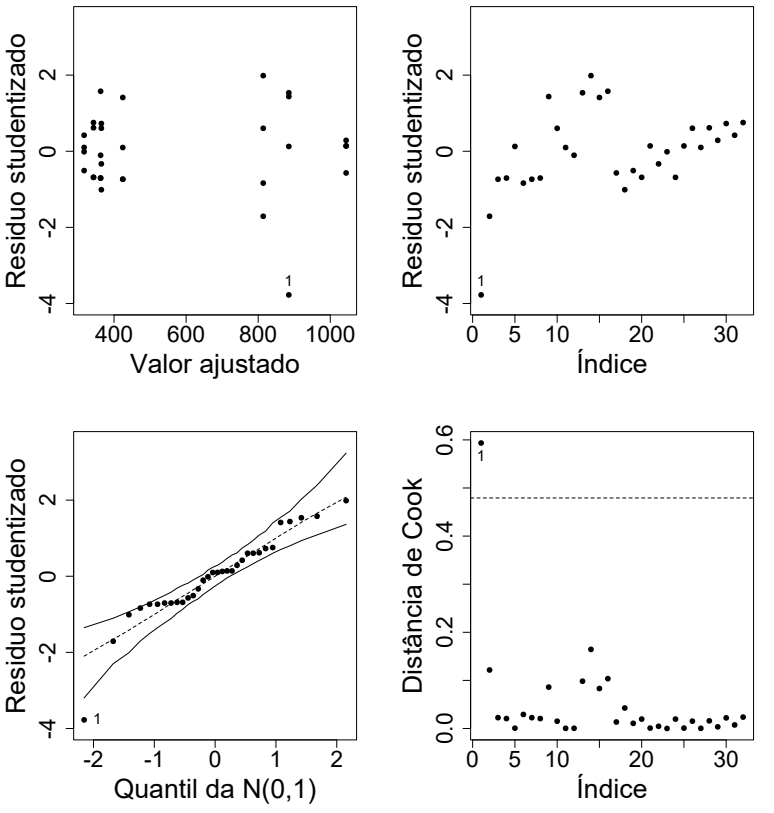

Figura 3 Análise de diagnóstico do modelo ajustado: Gráfico do resíduo studentizado contra valor ajustado (painel superior esquerdo); Gráfico do resíduo studentizado contra índice das observações (painel superior direito) e Gráfico normal de probabilidade do resíduo studentizado com banda de 95% de confiança (painel inferior esquerdo) e Gráfico da distância de Cook contra índice das observações (painel inferior direito).

A Figura 2 mostra o gráfico dos perfis ajustados do modelo com interação, note que este gráfico coincide com o gráfico dos perfis médios da análise descritiva uma vez que no modelo de análise de variância de dois fatores com interação o número de parâmetros é igual ao número de médias.

12. Considere o conjunto de dados `BigMac2003` disponível na biblioteca `alr4` (Weisberg, 2018) do R, em que são descritas as seguintes variáveis de 69 cidades de diversos países: `BigMac`: Minutos de trabalho para comprar um Big Mac; `Bread`: Minutos de trabalho para comprar 1kg de pão; `Rice`: Minutos de trabalho para comprar 1kg de arroz; `FoodIndex`: Índice de preços de alimentos; `Bus`: Valor da passagem de ônibus (em USD); `Apt`: Valor do aluguel de um apartamento padrão de três dormitórios (em USD); `TeachGI`: Salário bruto anual de um professor de ensino fundamental (em 1000 USD); `TeachNI`: Salário líquido anual de um professor de ensino fundamental (em 1000 USD); `TaxRate`: Imposto pago por um professor de ensino fundamental (em porcentagem) e `TeachHours`: Carga horária semanal de um professor de ensino fundamental (em horas).

    O objetivo principal do estudo é relacionar a variável `BigMac` com as demais variáveis explicativas. A fim de melhorarmos a aproximação de normalidade consideraramos $\texttt{lBigMac}_i = \log(\texttt{BigMac}_i)$ na i-ésima cidade como variável resposta. A Figura 4 apresenta os diagramas de dispersão (com tendências cúbicas) entre a variável resposta e cada uma das nove variáveis explicativas, vemos descritivamente que a relação de `lBigMac` tem forma quadrática com as variáveis `Bread`, `Rice`, `FoodIndex`, `Bus`, `Apt`, `TeachGI` e `TeachNI`, e tem forma linear com as variáveis `TaxRate` e `teachHours`.

    Padronizamos todas as variáveis explicativas afim de tornar os coeficientes diretamente comparáveis, segundo a seguinte padronização $\texttt{sBread}_i = (\texttt{Bread}_i - \texttt{md.Bread})/\texttt{sd.Bread}$, ilustrada para variável explicativa `Bread`, em que `md.Bread` e `sd.Bread` são respectivamente a média e o desvio padrão da variável. Assim, todas as variáveis têm média igual a 0 e variância igual 1. Ajustamos um modelo linear normal para `lBigMac` com todas as variáveis explicativas padronizadas e com os 7 efeitos quadráticos identificados na análise descritiva, através do algoritmo de seleção `stepAIC()`, da biblioteca `MASS` (Ripley, 2023), reduzimos ao seguinte modelo

    1. $\texttt{lBigMac}_i | \mathbf{x}_i \overset{\text{ind}}{\sim} \text{NO}(\mu_i, \sigma)$,
    2. $$\mu_i = \begin{cases} \alpha + \beta_1 \texttt{sBread}_i + \gamma_1 \texttt{sBread}_i^2 + \\ \quad \beta_2 \texttt{sRice}_i + \gamma_2 \texttt{sRice}_i^2 + \\ \quad \beta_3 \texttt{sFoodIndex}_i + \gamma_3 \texttt{sFoodIndex}_i^2 + \\ \quad \beta_4 \texttt{sTeachNI}_i + \gamma_4 \texttt{sTeachNI}_i^2 + \\ \quad \beta_5 \texttt{sTeachHours}_i, \end{cases}$$

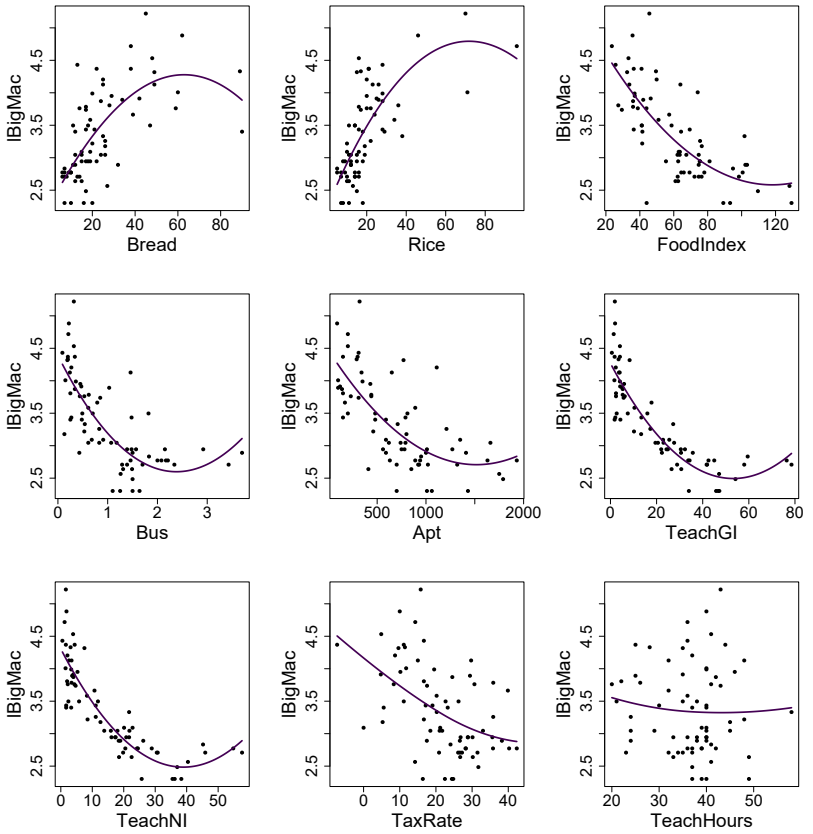

Figura 4 Diagramas de dispersão (com tendências cúbicas) entre a variável resposta e cada uma das nove variáveis explicativas.

Tabela 2 Estimativas dos parâmetros e erros padrões do modelo selecionado com AIC = 31.74 para relacionar `lBigMac` com as demais variáveis explicativas.

Parâmetro	Estimativa	Erro padrão	valor-P
α	3.36	0.03	< 0.0001
β_1	1.09	0.40	0.0082
γ_1	-0.75	0.33	0.0266
β_2	1.09	0.38	0.0054
γ_2	-0.19	0.33	0.5679
β_3	-1.46	0.56	0.0113
γ_3	-0.05	0.33	0.8900
β_4	-1.94	0.71	0.0085
γ_4	1.56	0.37	< 0.0001
β_5	0.41	0.30	0.1792
σ	0.28		

em que β_j e γ_k denotam os efeitos linear e quadrático da j-ésima variável explicativa para $i = 1, \ldots, 69$, $j = 1, \ldots, 5$ e $k = 1, \ldots, 4$. A Tabela 2 apresenta as estimativas dos parâmetros e erros padrões do modelo selecionado e a Figura 5 mostra os gráficos da análise de resíduos e de sensibilidade do modelo, verificamos que as suposições de homocedasticidade da variância, independência e normalidade das observações parecem satisfeitas. O gráfico da distância de Cook destaca as observações #14 e #57 que correspondem as cidades de Shanghi e Caracas com valores de `sBread` de 3.67 e 3.62, que são muito maiores que os demais. Como análise confirmatória, quando removemos tais observações o mesmo modelo é selecionado. Portanto, como não há mudança na seleção e as suposições estão satisfeitas a análise de diagnóstico validou o procedimento de seleção realizado.

13. Considere o conjunto de dados `motorins` disponível na biblioteca `faraway` (Faraway 2022) do R, em que são apresentadas informações relacionadas a 1797 grupos de apólices de seguro de automóvel no ano de 1977 na Suécia, em que são descritas as seguintes variáveis: `Zone`: 1, ..., 6 ou 7 para as sete regiões do país e `perd`: Valor pago por sinistro (em coroas suecas). Em particular, há interesse em saber se há diferenças significativas entre o seguro médio pago por sinistro em 7 regiões do país. A fim de obtermos uma melhor aproximação para a normalidade, vamos considerar a variável $\texttt{lperd}_{ij} = \log(\texttt{perd}_{ij})$ da i-ésima região e j-ésimo grupo de apólice como resposta. A Figura 6 apresenta os boxplots de `lperd` segundo a região, vemos descritivamente alguns outliers em todas as regiões e que apenas a região 7 parece ser diferente das demais.

    Seja $\texttt{lperd}_{ij}$ o valor pago por sinistro da i-ésima região e j-ésimo grupo de apólice, supondo o seguinte modelo de comparação de médias

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2),$$

    em que μ_i denota a média da i-ésima região para $i = 1, \ldots, 7$ e $j = 1, \ldots, n_i$. A Tabela 3 apresenta a tabela ANOVA do ajuste do modelo. Notamos que a hipótese de homogeneidade de médias entre as regiões é rejeitada, avaliada com nível de 5%, logo existe ao menos uma região com média diferente das demais. Note que como temos muitas observações dentro dos grupos, as suposições de homocedasticidade da variância, independência e normalidade das observações podem ser relaxadas e a decisão tomada a respeito da homogeneidade dos grupos ainda pode ser validada.

    Aplicando o método de Tukey através do comando `TukeyHSD()` para verificar quais das 21 possíveis diferenças entre as regiões são significativas, vemos na Figura 7 que apenas os intervalos (com 95% de confiança conjunta) das diferenças envolvendo a região 7 não contém o 0. Portanto, a região 7 é a única diferente das demais com média de seguro pago por sinistro menor que as outras.

Tabela 3 Tabela ANOVA do modelo para ajustar o seguro médio pago por sinistro nas regiões do país.

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	valor-F	valor-P
`Zone`	6	19.42	3.2367	6.5796	< 0.0001
Total	1790	880.56	0.4919		

14. Considere o conjunto de dados `fuel2001` disponível na biblioteca `alr4` (Weisberg, 2018) do R, em que são descritas as seguintes variáveis referentes a 50 estados norte-americanos mais o Distrito de Columbia no ano

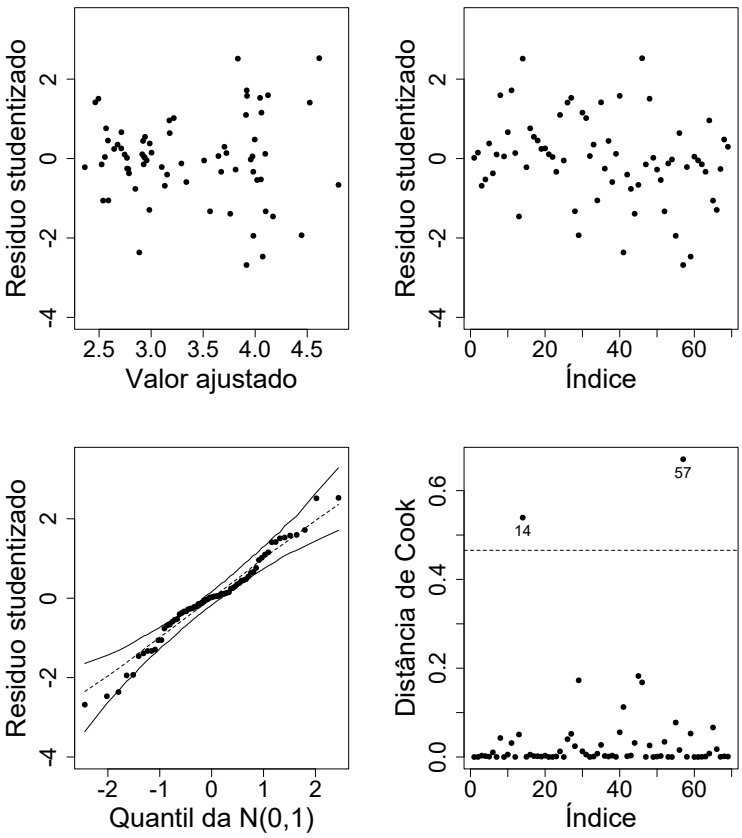

Figura 5 Análise de diagnóstico do modelo selecionado: Gráfico do resíduo studentizado contra valor ajustado (painel superior esquerdo); Gráfico do resíduo studentizado contra índice das observações (painel superior direito) e Gráfico normal de probabilidade do resíduo studentizado com banda de 95% de confiança (painel inferior esquerdo) e Gráfico da distância de Cook contra índice das observações (painel inferior direito).

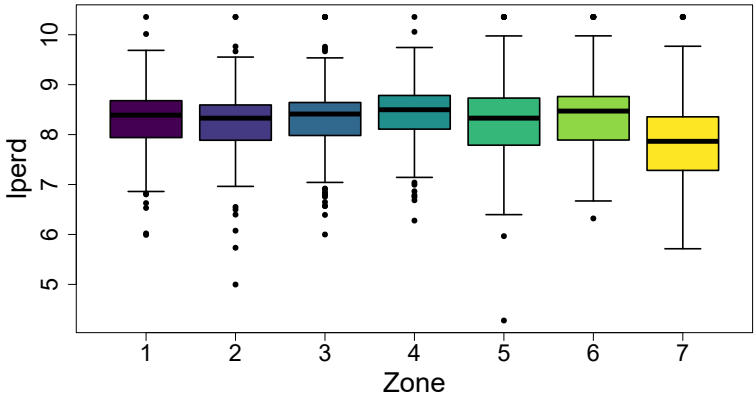

Figura 6 Boxplots de `lperd` segundo a região.

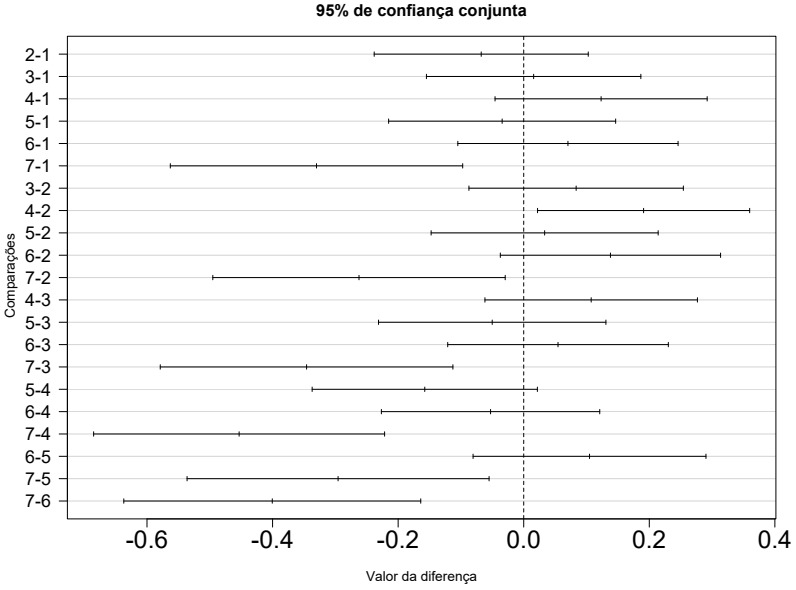

Figura 7 Intervalos para as diferenças entre as regiões com de 95% de confiança conjunta.

de 2001: `UF`: Unidade da federação; `Drivers`: Número de motoristas licenciados; `FuelC`: Total de gasolina vendida (em mil galões); `Income`: Renda per capita em 2000 (em mil USD); `Miles`: Total de milhas em estradas federais; `MPC`: Milhas per capita percorridas; `Pop`: População com 16 anos ou mais e `Tax`: Preço da gasolina (em cents/galão).

A fim de possibilitar uma comparação entre as UFs duas novas variáveis foram consideradas $\texttt{Fuel}_i = 1000\texttt{FuelC}_i/\texttt{Pop}_i$ e $\texttt{Dlic}_i =$

$1000\texttt{Drivers}_i/\texttt{Pop}_i$, além da variável $\texttt{lMiles}_i = \log(\texttt{Miles})_i$, para a i-ésima unidade da federação. Consideramos como resposta a variável `Fuel` e como variáveis explicativas: `Dlic`, `lMiles`, `Income` e `Tax`. A Figura 8 mostra o boxplot robusto (gerado com o comando `adjbox()` da biblioteca `robustbase` (Maechler, 2023)) e a densidade empírica para a variável resposta. A Figura 9 apresenta os diagramas de dispersão (com tendências cúbicas) entre cada variável explicativa e a variável resposta, notamos descritivamente que a distribuição marginal da variável resposta `Fuel` parece ser simétrica, tendo relação de forma não linear com `Dlic` e `lMiles` e de forma linear com `Income` e `Tax`.

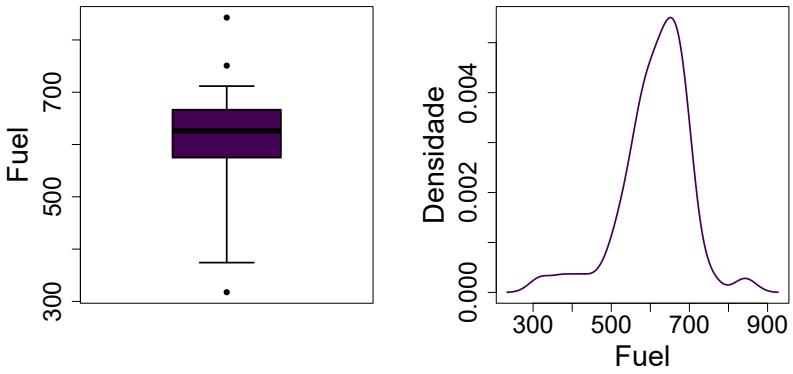

Figura 8 Boxplot robusto (painel esquerdo) e a densidade empírica (painel direito) para a variável resposta.

Ajustamos um modelo linear normal para `Fuel` com todas as variáveis explicativas consideradas e somente com efeitos lineares, através do algoritmo de seleção `stepAIC()` da biblioteca `MASS` (Ripley, 2023) verificamos que o modelo não é reduzido. Tentamos incluir, um a um, cada um dos 6 efeitos de interação e avaliando o teste da inclusão com um nível de significância de 5% obtivemos o seguinte modelo

1. $\texttt{Fuel}_i|\mathbf{x}_i \overset{\text{ind}}{\sim} \text{NO}(\mu_i, \sigma)$,
2. $\mu_i = \begin{cases} \alpha + \beta_1\texttt{Dlic}_i + \beta_2\texttt{lMiles}_i + \beta_3\texttt{Income}_i + \beta_4\texttt{Tax}_i + \\ \quad + \gamma\texttt{lMiles}_i\texttt{Tax}_i, \end{cases}$

em que β_j denota o efeito linear da j-ésima variável explicativa para $i = 1, \ldots, 51$ e $j = 1, \ldots, 4$. A Tabela 4 apresenta as estimativas dos parâmetros e erros padrões do modelo selecionado.

A Figura 10 mostra os gráficos da análise de resíduos e de sensibilidade do modelo selecionado, verificamos que as suposições de homocedasticidade da variância, independência e normalidade das observações parecem satisfeitas. Os gráficos de diagnóstico destacam as observações #51, #2,

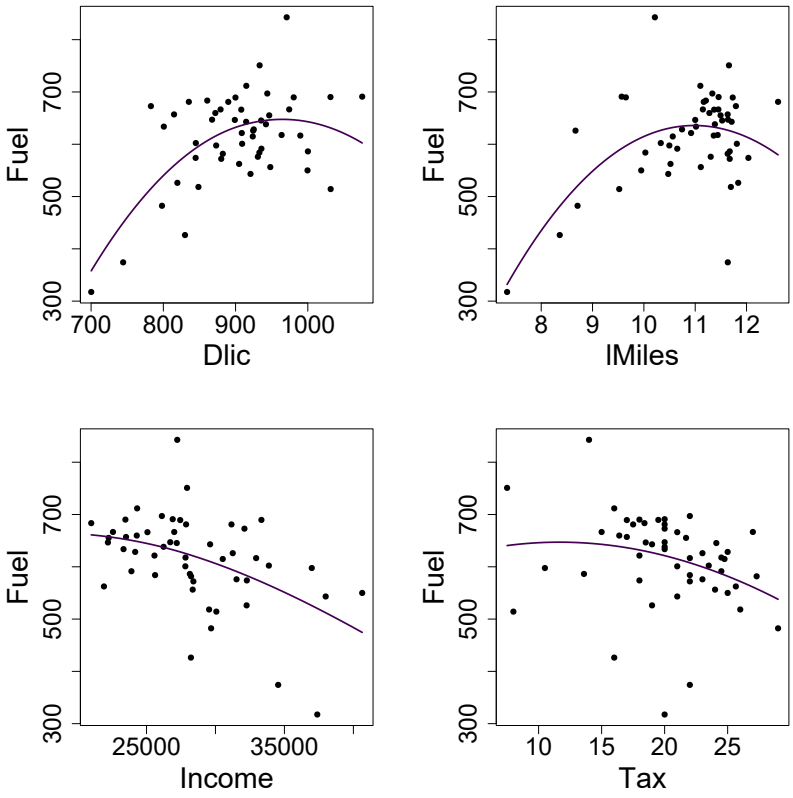

Figura 9 Diagramas de dispersão (com tendências cúbicas) entre a variável resposta e cada uma das quatro variáveis explicativas consideradas.

#9, #40 e #10 que correspondem as unidades da federação de WY, AK, DC, RI e FL. Como análise confirmatória, a Tabela 5 apresenta as estimativas dos parâmetros e erros padrões do modelo selecionado quando removemos tais observações, vemos que apesar de mudanças numéricas

Tabela 4 Estimativas dos parâmetros e erros padrões do modelo selecionado com AIC = 572.99 para relacionar `Fuel` com as demais variáveis explicativas consideradas.

Parâmetro	Estimativa	Erro padrão	valor-P
α	-877.018	469.654	0.0684
β_1	0.512	0.124	0.0001
β_2	119.603	39.848	0.0044
β_3	-0.006	0.002	0.0057
β_4	45.069	20.714	0.0349
γ_1	-4.616	1.931	0.0211
σ	61.80		

dos estimadores nenhum sinal ou inferência são alterados. Portanto, como não há mudanças de decisões e as suposições estão satisfeitas a análise de diagnóstico validou o procedimento de seleção realizado e o modelo escolhido.

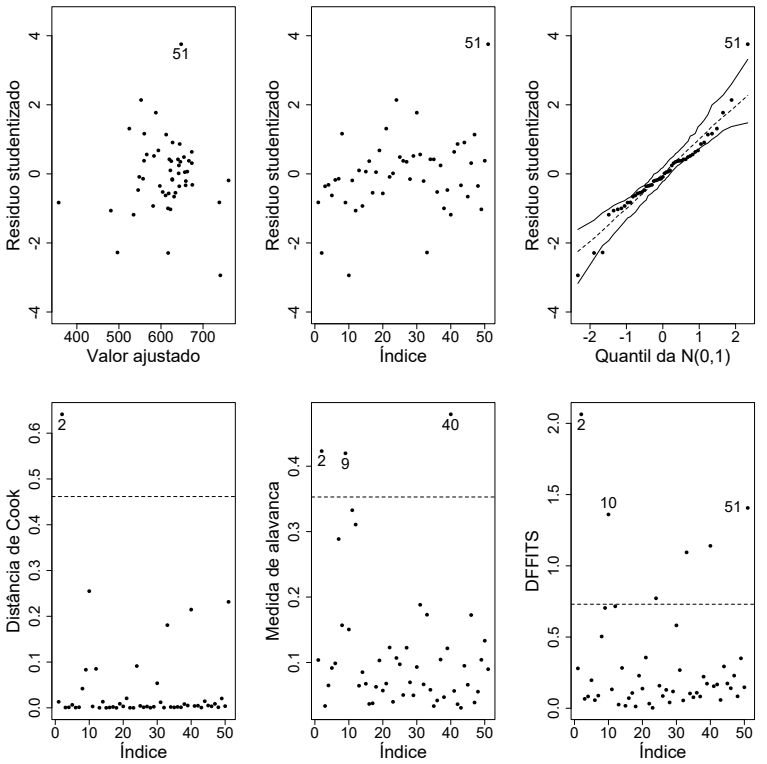

Figura 10 Análise de diagnóstico do modelo selecionado: Gráfico do resíduo studentizado contra valor ajustado (painel superior esquerdo); Gráfico do resíduo studentizado contra índice das observações (painel superior central); Gráfico normal de probabilidade do resíduo studentizado com banda de 95% de confiança (painel superior direito); Gráfico da distância de Cook contra índice das observações (painel inferior esquerdo); Gráfico da medida de alavanca contra índice das observações (painel inferior central) e Gráfico da distância DFFITS contra índice das observações (painel inferior esquerdo).

Quanto as interpretações, temos que o aumento de 10 unidades na quantidade de motoristas por habitantes (`Dlic`) de uma unidade da federação causa aumento de 5.12 na média do total de gasolina vendida por habitante (`Fuel`) da unidade da federação e o aumento de 100 unidades na renda por habitante (`Income`) causa diminuição de 6 unidades na média de `Fuel`, em todos os casos, mantendo as demais variáveis constantes. Como existe efeito de interação entre `lMiles` e `Tax`, não podemos fazer

Tabela 5 Estimativas dos parâmetros e erros padrões do modelo selecionado com AIC = 493.60 para relacionar `Fuel` com as demais variáveis explicativas consideradas removendo as observações destacados na análise de diagnóstico.

Parâmetro	Estimativa	Erro padrão	valor-P
α	-1267.69	555.541	0.0279
β_1	0.492	0.123	0.0003
β_2	156.873	50.604	0.0035
β_3	-0.005	0.002	0.0019
β_4	67.972	28.234	0.0208
γ_1	-6.768	2.537	0.0110
σ	47.65		

interpretações como as anteriores. Assim temos na Figura 11 apresenta uma visualização gráfica da média ajustada de `Fuel` em função de `lMiles` e `Tax` para `Dlic` = 903.68 e `Income` = 28403.90 fixados.

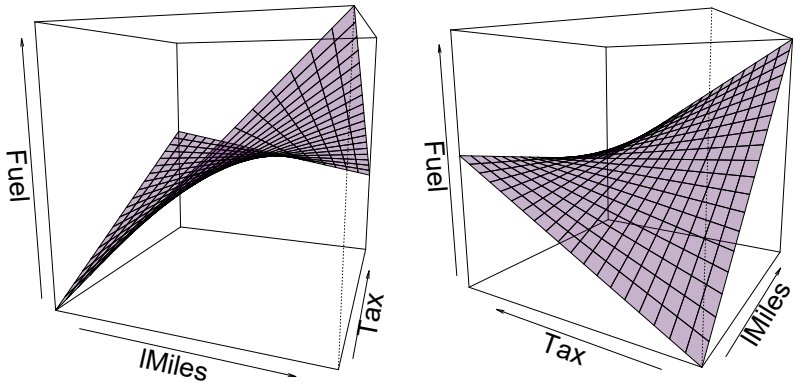

Figura 11 Visualização gráfica da média ajustada de `Fuel` em função de `lMiles` e `Tax` para `Dlic` = 903.68 e `Income` = 28403.90 fixados.

Referências

Faraway J (2022) faraway: Functions and Datasets for Books by Julian Faraway. R package version 1.0.8. https://cran.r-project.org/package=faraway

Lawless JF (1982) Statistical Models and Methods for Lifetime Data. Wiley

Maechler M (2023) robustbase: Basic Robust Statistics. R package version 0.95-1. https://cran.r-project.org/package=robustbase

Paula GA (2023) Regressão linear múltipla (versão parcial preliminar). Notas de aula atualizadas em 03-23 do curso MAE0328

Ripley B (2023) MASS: Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-60. https://cran.r-project.org/package=MASS

Weisberg S (2018) alr4: Data to Accompany Applied Linear Regression 4th Edition. R package version 1.0.6. https://cran.r-project.org/package=alr4