



COMPORTAMENTO DE PREÇO DE CARROS IMPORTADOS, UMA APLICAÇÃO DA REGRESSÃO LINEAR MÚLTIPLA

PRICE BEHAVIOR OF IMPORTED CARS, AN APPLICATION OF MULTIPLE LINEAR REGRESSION

Gabriel Penha*,

Moisés Augusto†,

Marla Lorrani‡

RESUMO

O preço de produtos ofertados é um objeto de interesse recorrente. Não é diferente na indústria automobilística, setor que chama atenção tanto de consumidores, quanto da academia. Neste trabalho, objetivou-se ajustar um modelo de regressão linear múltiplo para o preço de carros importados, presentes na base de dados *import-85*. Para tal, consideraram-se os pressupostos do modelo, como normalidade, homoscedasticidade e independência, medidas de qualidade do ajuste, como os coeficiente de determinação e o coeficiente de determinação ajustado e de seleção de variáveis, como o *Stepwise*. Para avaliar os ajustes entre o Modelo 1 e o Modelo final, o critério de informação de Akaike e o Bayesiano especialmente importantes, sendo que destes, o BIC recebeu uma maior importância do ponto de vista dos autores, que consideram que ele seria mais indicado para lidar com multicolinearidade, sem que técnicas de análise mais complexas fossem empregadas. Em geral, as suposições foram verificadas e o ajuste foi considerado satisfatório, apesar de ressalvas envolvendo a independência dos resíduos no modelo final.

Palavras-chave: Modelo. Regressão. Carros importados. Preço. Pressupostos.

1 INTRODUÇÃO

A indústria automobilística é um setor da economia que chama muita atenção, tanto de eventuais consumidores, quanto da academia. (FERREIRA; RIBEIRO, 2003; GONZALEZ; MARTINS, 2011, 2007)


Assim como em diversos mercados do sistema econômico vigente, o preço dos produtos ofertados é um objeto de interesse recorrente. Não a toa, é possível encontrar inúmeras bases de dados em plataformas como o *Kaggle*, cujo objetivo, é justamente prever o preço de um produto com base em suas características.


Nesta mesma linha, a base de dados nomeada de *import-85*, disponibilizada em Dua e Graff (2017) disponibiliza informações sobre diversas características de 205 carros importados - em relação aos Estados Unidos -, como, por exemplo, o fabricante do automóvel, o tipo de combustível com que ele opera e seu número de portas.


Ao todo, são 26 características - variáveis -, sendo 25 delas explicativas em relação ao preço, dado em dólares (\$).

Assim sendo, este trabalho objetivou ajustar um modelo de regressão linear múltiplo, na tentativa de explicar o preço dos carros, levando em conta os diversos pressupostos que este método carrega, como a normalidade, a heterocedasticidade e a independência dos resíduos, além da multicolinearidade.

A abordagem da regressão linear múltipla para ajustes e previsões em bases de dados com respostas quantitativas é utilizada na literatura, seja para definição de um modelo em si, ou para comparação com outros possíveis métodos. (DING et al., 2014; RESENDE et al., 1996)

*  Departamento de Estatística, UFBA, Bacharelado em Estatística; penha.gabriel@ufba.br.

†  Departamento de Estatística, UFBA, Bacharelado em Estatística; moises.augusto@ufba.br.

‡  Departamento de Estatística, UFBA, Bacharelado em Estatística; marla.lorrani@ufba.br.

Neste trabalho, outros métodos não foram empregados objetivando uma comparação com o modelo linear; o que acaba sendo uma interessante abordagem para trabalhos futuros com a mesma base de dados.

Além disso, não se utilizaram técnicas mais sofisticadas que as comumente vistas em cursos de análise de regressão focados no método linear e, assim sendo, problemas como a multicolinearidade e a independência dos resíduos precisaram ser resolvidos com base na seleção de variáveis.

O presente trabalho está subdividido em seções; na seção seguinte, ter-se-á a metodologia, em que as técnicas estatísticas empregadas, as suposições para elas e um maior detalhamento da base de dados utilizada serão brevemente explicitados. Em seguida, a seção de resultados e discussão trará algumas nuances da análise exploratória dos dados e apresentará dois modelos: o Modelo 1, (que também será chamado de base) e o Modelo final, definido após a seleção de variáveis através do *Stepwise* e eventuais ajustes, com base na análise de resíduos. Finalmente, a conclusão irá ressaltar algumas das considerações importantes levantadas durante as análises, além de elencar eventuais caminhos posteriores que podem ser tomados a partir desta análise.

2 METODOLOGIA

O ajuste de um modelo de regressão linear perpassa por algumas análises não necessariamente exclusivas dessa metodologia. Assim como em outros procedimentos da estatística, é desejável que se compreenda o comportamento geral dos dados por meio de análises exploratórias, sejam elas analíticas ou visuais.

Assim sendo, além da visualização de distribuições e relacionamento das variáveis, como abordado em Demétrio e Zocchi (2006), uma forma de verificar o relacionamento linear de variáveis quantitativas é o coeficiente de correlação de Pearson.

Dadas duas variáveis X, Y , a correlação entre elas é um valor que varia entre -1 e 1 ; sendo que valores negativos indicam uma associação inversa (isto é, enquanto os valores observados em uma variável crescem, os valores observados na outra tendem a decrescer), e valores próximos de 0 indicam uma falta de relação. Esta medida é definida assim:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}. \quad (1)$$

Sendo ρ o coeficiente mencionado.

Vale dizer que, por vezes, apesar de X e Y não parecerem ter uma associação linear forte diretamente, acontece de uma das variáveis se associar bem com uma função polinomial da outra (Y bem associado com X^2 , por exemplo). É uma situação recorrente que, inclusive, ocorreu com algumas das preditoras consideradas aqui, e o preço.

Dito isso, é importante notar, que como mencionado, o coeficiente de Pearson serve para averiguar o relacionamento de variáveis **quantitativas**. Além de poder ser utilizado para verificar a força da relação de uma preditora com o preço, pode dar indícios de multicolinearidade entre esse tipo de variável; no entanto, não ajuda muito com variáveis qualitativas.

Para este caso, em Bécue-Bertaut e Pagès (2008) o coeficiente V de Crammer é utilizado. Em linhas gerais, ele serve para avaliar o relacionamento entre variáveis qualitativas - o que serviu para dar indícios de multicolinearidade entre algumas das preditoras do conjunto de dados -. Para este coeficiente, valores próximos de 1 estão muito associados, enquanto valores próximos de 0 não estão.

Em modo geral, nas análises aqui realizadas, desejava-se que o relacionamento entre as preditoras qualitativas e o preço, mensuradas pelo coeficiente de Pearson, fossem, em módulo, próximas de 1 , enquanto que o relacionamento das preditoras quantitativas (entre si) e das qualitativas (também entre si), fossem próximas de 0 .

Para a análise de modelos múltiplos, de modo geral, alguns pressupostos são tomados, inicialmente, como verdade e, posteriormente, precisam ser checados.

Espera-se, que:

1. A relação entre as preditoras e a resposta se dê de uma forma linear; ou seja, para X , uma matriz de co-variáveis, o modelo seja da forma $Y = \beta X + \epsilon$, em que Y, β e ϵ são vetores com dimensões apropriadas. (HOFFMANN; VIEIRA, 1998)
2. Espera-se que os erros sejam independentes, normais, com média 0 e variância constante, isto é: $E(\epsilon) = 0, Var(\epsilon) = I\sigma^2, Cov(\epsilon_i, \epsilon_j) = 0, \forall i, j = 1, \dots, n$ e $\epsilon_i \sim N(0, \sigma^2)$.
3. E, conseqüentemente, que $Y|X \sim N(X\beta, I\sigma^2)$. (DEMÉTRIO; ZOCCHI, 2006)

Os pressupostos expressos em (2), na lista acima, costumeiramente são verificados através de análises gráficas e testes de hipóteses.

Aqui, o pressuposto de normalidade foi avaliado com o teste Shapiro-Wilk, que supõe, sob hipótese nula, que os resíduos do modelo são normais; o que, quando verdade, indica que os erros também o são. A estatística do teste de Shapiro-Wilk é dada por $W = \frac{b^2}{\sum (X_i - \bar{X})^2}$. (RAWLINGS; PANTULA; DICKEY, 2001)

A suposição de homoscedasticidade (variância dos erros é constante) também é checada através dos resíduos; o teste utilizado neste trabalho foi o de Breush-Pagan. Para verificar a estatística de teste utilizada em tal procedimento, o leitor é encaminhado para (DEMÉTRIO; ZOCCHI, 2006);

O teste de Durbin Watson verifica a independência dos resíduos, sua estatística é dada da seguinte maneira: $d = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}$. Uma regra prática, é que se $1,5 \leq d \leq 2,5$, os resíduos são independentes. (RAWLINGS; PANTULA; DICKEY, 2001)

Além disso, avaliou-se a multicolinearidade das co-variáveis através do fator de inflação da variância (*Variance inflation factor - VIFs*); definidos por: $VIF_j = \frac{1}{1-R_j^2}$. Altos valores de *VIFs* indicam uma maior multicolinearidade; uma regra prática é considerar como não problemáticas variáveis com *VIFs* inferiores a 10.

O método de seleção de variáveis *Stepwise* foi implementado para que o modelo final pudesse ser escolhido a partir deste; o procedimento levou em conta métricas como *AIC* - critério de informação de Akaike -, definido por: $-n \log(\frac{SQRes}{n}) + 2k$; *BIC* - critério de informação Bayesiano -, definido assim: $BIC = -n \log(\frac{SQRes}{n}) + k \log(n)$; R^2 , o coeficiente de determinação (R^2), que corresponde à proporção da variabilidade dos dados explicada pela regressão: $R^2 = \frac{SQReg}{SQTotal}$; coeficiente de determinação ajustado $R_\alpha^2 = R_\alpha^2 = R^2 - \frac{1}{n-2} * (1 - R^2)$; além do coeficiente de C_p de Mallows, $C_p = \frac{1}{\sigma^2} * \sum E[\hat{Y}_i - E(Y_i)]^2$. (DEMÉTRIO; ZOCCHI, 2006; VITTINGHOFF et al., 2006).

Ao final do método, o melhor modelo teve seus pressupostos novamente verificados; alguns dos problemas se mantinham e, dessa forma, precisou-se realizar uma transformação Box-Cox. Essa transformação é aplicável quando se deseja diminuir a heterocedasticidade do modelo; e, por vezes, pode ajudar a melhorar outros pressupostos como a normalidade. (RAWLINGS; PANTULA; DICKEY, 2001)

A sua expressão é definida neste documento, mais adiante, quando o contexto de sua aplicação nos dados foi explicada.

3 RESULTADOS E DISCUSSÃO

Considerando a metodologia de análise mencionada na seção anterior, é conveniente reportar os resultados obtidos de duas diferentes maneiras - em duas diferentes subseções -: resultados exploratórios e modelagem. A primeira das subseções trará alguns dos pontos mais importantes da análise exploratória realizada, enquanto a segunda apresentará o Modelo 1 e o Modelo final, além de alguns detalhes de suas respectivas análises de resíduo.

3.1 Resultados exploratórios

Os resultados exploratórios serão apresentados através da análise descritiva da variável resposta preço, de alguns poucos gráficos bi-variados de variáveis qualitativas para com o preço, e dos diagramas de dispersão bi-variados considerando as variáveis quantitativas, possível graças ao pacote *GGally* do R.

Na Tabela 1 é possível visualizar as estatísticas descritivas do preço dos automóveis.

Tabela 1 – Descritivas - Preço	
Estatística	Valor (\$)
Mínimo	5118,00
1º Quartil	7738,00
Mediana	10245,00
Média	13285,00
3º Quartil	16515,00
Máximo	45400,00

Fonte: Autores (2021)

Como é possível perceber, a média é \$3000 mais cara que a mediana, o que indica uma assimetria à direita (ou positiva).

Na Figura 1, é possível visualizar os gráficos de dispersão entre as variáveis quantitativas, incluindo o preço, e os coeficientes de correlação de Pearson calculados a cada duas delas. Salienta-se que, a variável taxa de compressão, já havia sido transformada (e, portanto, era qualitativa).

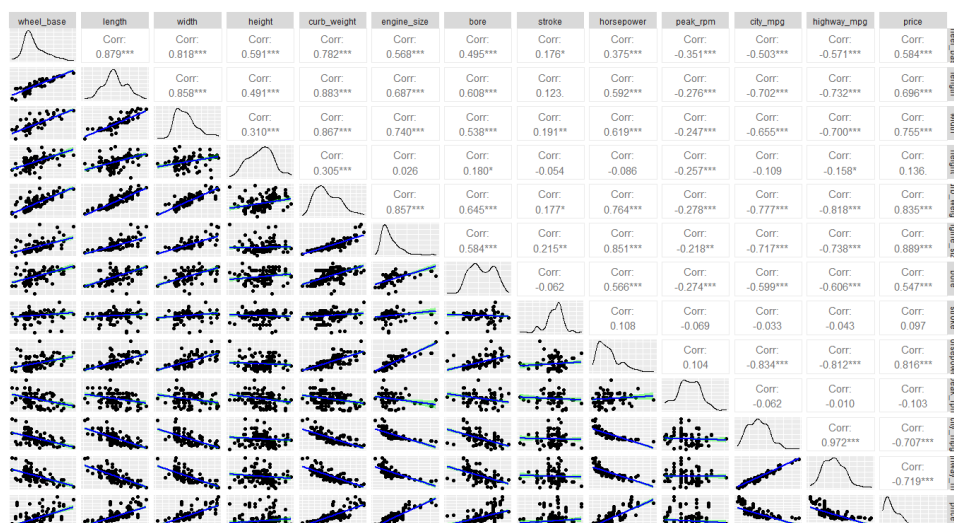


Figura 1 – Relacionamento entre variáveis quantitativas

Fonte: Autores (2021)

Como é possível ver, várias das preditoras pareciam se relacionar bem com o preço; entretanto, elas

também relacionavam-se bem entre si. Apesar desse não ser o melhor indicador existente de multicolinearidade, tratava-se de um forte indício de que este seria um dos problemas no ajuste do modelo linear. Em especial, as variáveis *city_mpg* e *highway_mpg*, que denotam, respectivamente, os registros de milhas por galão para cada carro nas cidades e nas rodovias, obtiveram uma correlação de 0,972, mais alta, inclusive, que o coeficiente de ambas as variáveis para com o preço. Apesar disso, vale ressaltar que, posteriormente, percebeu-se que a relação entre essas duas variáveis e o preço não se dava de modo linear, mas de modo polinomial. Isto, no entanto, não diminuiu as preocupações com a multicolinearidade.

A seguir, será possível ver alguns dos gráficos bi-variados entre as variáveis qualitativas e o preço.

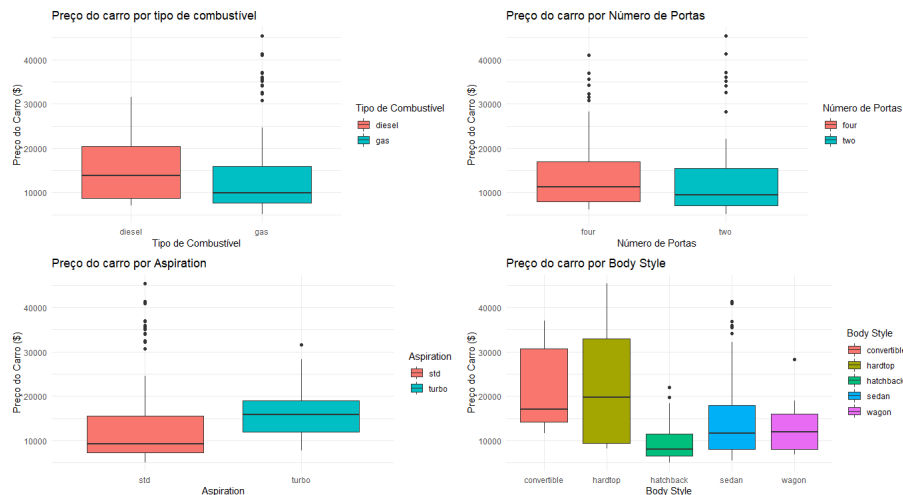


Figura 2 – Aspiração; Número de portas; Tipo de combustível; Estilo de corpo do carro Vs. Preço
 Fonte: Autores (2021)

Como é possível ver na Figura 2, o número de portas não parecia ter uma significância considerável (ao menos individualmente) para com o preço; o que, no entanto, difere do restante das variáveis.

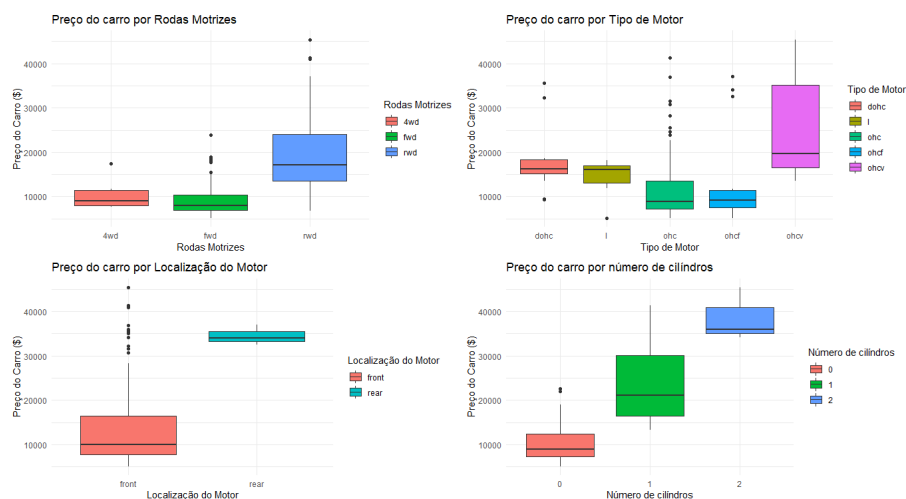


Figura 3 – Rodas motrizes; Tipo de motor; Localização do motor; Número de cilindros Vs. Preço
 Fonte: Autores (2021)

A Figura 3, apresenta a relação entre mais algumas variáveis qualitativas para com o preço. Aqui, todas as variáveis parecem ser significantes em relação ao preço, em especial a localização do motor e o número de cilindros. Vale ressaltar que a esse ponto, a preditora número de cilindros já havia sido recategorizada; em

geral, os carros que tinham mais de oito cilindros receberam a categoria 2, os que tinham menos que quatro, receberam 0 e os demais receberam 1.

Para finalizar a subseção que trata sobre a análise exploratória, verificaram-se, através de gráficos, possíveis interações entre as variáveis qualitativas e as quantitativas.

Na Figura 4 é possível visualizar um exemplo desses gráficos. As variáveis consideradas nela foram *city_mpg*, *aspiration* e *price*, que denotam de modo respectivo o número de milhas por galão que o carro faz na cidade, sua aspiração e preço.

Neste gráfico é possível perceber indícios de que há uma interação entre as duas variáveis preditoras na explicação do preço.

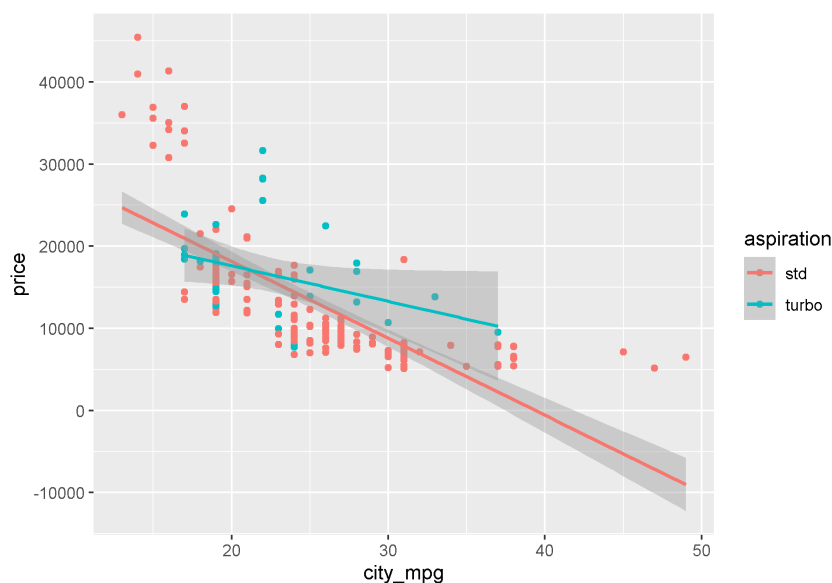


Figura 4 – Milhas por galão na cidade Vs. Aspiração Vs. Preço

Fonte: Autores (2021)

Vale salientar, no entanto, que por conta de problemas com multicolinearidade (que acarretaram em uma simplificação do modelo final), em conjunto com problemas técnicos na avaliação de termos de interação por métodos como o *Stepwise*, o uso de tais termos foi evitado na modelagem.

Para maiores detalhes sobre a exploratória, incluindo o acesso aos códigos, outros gráficos e procedimentos utilizados, o leitor é encaminhado ao endereço disponibilizado na seção de anexos, ao final do documento.

3.2 Modelagem

Como já mencionado em versões anteriores deste documento (ver endereço *web* nos anexos, na pasta documentos), alguns modelos lineares simples foram ajustados, para verificar o relacionamento entre algumas das preditoras quantitativas e a variável resposta. A partir disso, ajustaram-se modelos polinomiais dessas preditoras, de modo que foi possível verifica em até que grau (se quadrático ou cúbico, por exemplo), elas mantinham uma significância em sua relação com o preço.

Essas informações foram utilizadas na construção dos modelos múltiplos pré *Stepwise*. O primeiro deles pode ser observado na Tabela 2.

Tabela 2 – Modelo 1

Parâmetro	Coefficiente	Erro Padrão	Valor - p
Intercepto	14652,3	1745,4	0,00
Comprimento	17930,7	8027,1	0,03
Comprimento ²	8130,2	4152,2	0,05
Largura	9305,1	6227,6	0,13
Largura ²	4906,1	3654,1	0,18
Peso de meio fio	19257,6	12568,9	0,13
Peso de meio fio ²	6776,1	5321,0	0,20
Tamanho do motor	1016,8	11313,6	0,93
Cavalos de força	11652,3	11421,4	0,31
MPG na cidade	-48640,0	13951,8	0,00
MPG na cidade ²	28057,2	9504,8	0,00
MPG na cidade ³	-11328,4	8705,1	0,19
MPG na cidade ⁴	7765,3	3210,8	0,01
MPG na rodovia	27258,3	12939,6	0,03
MPG na rodovia ²	-14561,8	9280,2	0,12
MPG na rodovia ³	1723,1	10362,6	0,87
Taxa de compressão (1)	2832,0	1235,9	0,02
Número de cilindros (1)	3321,2	896,1	0,00
Número de cilindros (2)	9777,3	3015,1	0,00
Tipo de motor (1)	-1292,1	1330,9	0,33
Tipo de motor (ohc)	2673,6	905,4	0,00
Tipo de motor (ohcf)	3001,5	1229,1	0,02
Tipo de motor (ohcv)	-2464,5	1209,5	0,04
Localização do motor (traseira)	8535,6	2254,4	0,00
Rodas motrizes (fwd)	-1245,1	1115,1	0,27
Rodas motrizes (rwd)	1297,3	1181,7	0,27
Estilo de corpo (hardtop)	-4230,8	1352,9	0,00
Estilo de corpo (hatchback)	4545,7	1156,6	0,00
Estilo de corpo (sedan)	-4210,9	1213,1	0,00
Estilo de corpo (wagon)	-5566,7	1321,9	0,00
Aspiração (turbo)	266,6	818,1	0,74
R ²	0,9354	R ² _α	0,9233
Estatística F	77,66 em 30 e 161 G.L	Valor-p:	≈ 0

Fonte: Autores (2021)

Para interpretar este modelo, tenha em mente que para as variáveis quantitativas (todas desde comprimento até taxa de compressão, sem incluir esta última) afetam a variável resposta, **em média**, no valor correspondente em coeficiente; considerando que as demais variáveis estejam fixadas. Exemplificando: Considerando as outras variáveis fixadas, cavalos de força afeta o preço, em média, em \$11652,3 por unidade; isto é, a cada acréscimo em unidades de cavalos de força, este é o valor esperado de acréscimo em preço.

Os parâmetros que estão ao lado de expoentes representam variáveis em que o ajuste foi polinomial; note que para este primeiro modelo, a variável *city_mpg* que, na tabela, é representada por “MPG na cidade” está elevada na quarta potência. A inclusão das variáveis de menor grau se trata de uma necessidade teórica do modelo linear.

Quanto às variáveis qualitativas (aquelas a partir de taxa de compressão), a respectiva categoria desta variável é representada entre parênteses; portanto, 2832,0 é o coeficiente que indica o efeito médio da taxa de compressão quando seu valor é a categoria (1), indicada na metodologia. Este efeito é calculado em comparação à categoria de taxa de compressão que não está presente no modelo, ou seja, a categoria (0).

Assim sendo, fixadas as demais variáveis, a taxa de compressão (1) afeta o preço, em média, em \$2832,00, positivamente, se comparada à taxa de compressão (0). As análises para as demais variáveis são análogas. Os erros padrão foram calculados com base na teoria de modelagem linear múltipla, um exercício interessante seria calcular tais valores de modo empírico, utilizando, por exemplo, o *bootstrap*, como descrito em James et al. (2013). Os valores-p foram calculados para indicar a significância de determinada variável (e categoria) em sua inclusão no modelo. O ideal (para que a variável seja significativa) é que este valor aproxime-se de zero. Neste caso, variáveis como o tamanho do motor (com valor-p próximo de 1, o máximo possível) provavelmente não são interessantes de serem mantidas no modelo. Vale dizer que para que variáveis qualitativas com uma ou mais categorias significantes permanecessem no modelo, por necessidades também teóricas, a menos que alguma recategorização bem embasada seja feita, as categorias não significantes também precisam permanecer no modelo. O mesmo é válido para variáveis polinomiais, como MPG na cidade. Para que a variável de grau 4 seja mantida, é preciso manter também a de grau 3.

Este modelo obteve coeficiente de determinação ajustado R^2_{α} de 0,9233, consideravelmente alto. Além disso, como pode-se observar no valor-p da estatística F (que testa a significância do modelo como um todo) na base da tabela, o ajuste é significativo.

Uma análise de resíduos mais rebuscada foi realizada para o Modelo 1. Aqui, no entanto, somente uma tabela resumindo as informações dos testes numéricos será exibida; de modo geral, eles representaram bem o que foi apurado na análise visual.

Tabela 3 – Análise de resíduos - Modelo 1

Pressuposto	Teste	Estatística	Valor-p
Normalidade	Shapiro-Wilk	0.9345	0
Homoscedasticidade	Breush-Pagan	75,207	0
Independência	Durbin Watson	1,4791	0
Multicolinearidade	Média dos VIFs	724,12	-

Fonte: Autores (2021)

Na Tabela 3, é possível ver uma sumarização de estatísticas calculadas para testar os pressupostos do modelo linear. Idealmente, para que os pressupostos sejam garantidos, espera-se que o valor-p seja grande, mais próximo de 1. Como é possível perceber, isso não ocorreu nem para a normalidade, nem para a homoscedasticidade, nem para a independência. No entanto, existem algumas ressalvas aqui; para o caso da independência, existe uma regra prática de que se a estatística de teste estiver entre 1,5 e 2,5, pode-se adotar independência; não é o caso aqui, mas o valor é suficientemente próximo. Quanto à multicolinearidade, a média dos VIFs (*Variance inflation factors*) não é exatamente um teste, mas se trata de uma regra prática, assim como os VIFs em si. Espera-se que os VIFs de cada uma das variáveis não supere 10 (novamente, uma regra prática), como é possível observar pela média desses valores, algumas coisas deram um pouco errado. As variáveis relativas à economia de combustível do carro (que possuem MPG em seu nome), tiveram VIFs superiores a 10000 em seus termos polinomiais, indício fortíssimo de multicolinearidade.

Diante de todos esses problemas, alguns pontos foram percebidos:

- Seria necessário realizar uma seleção de variáveis, que foi realizada com o *Stepwise*, para ambas as direções (adicionando e retirando variáveis), sendo que estas eram acrescentadas, ou não, com base em quão significativa se mostrava para explicar a resposta;
- Possivelmente, mesmo após a seleção de variáveis seria necessário realizar uma filtragem, em geral os valores dos VIFs foram muito altos; e como técnicas mais sofisticadas para lidar com tais problemas não seriam consideradas, este seria o caminho indicado;

- Além disso, em virtude do notado em relação à normalidade e à homoscedasticidade, o uso de transformações, como a Box-Cox, seriam (e acabaram, de fato, sendo), uma abordagem interessante.
- Dito isso, alguns pontos da base de dados apresentaram valores extremamente inconsistentes (e, em momentos, acabavam por impedir os ajustes por pertencerem à categorias em que só eles se faziam presentes), alguns destes foram retirados (identificados pelos números 67 e 48) e outros tiveram suas variáveis recategorizadas, o que evitou seu descarte.

Considerando todos pontos mencionados, para a escolha do modelo final com auxílio do *Stepwise* avaliaram-se critérios como AIC (critério de informação de Akaike), BIC (critério de informação Bayesiano), R^2 (coeficiente de determinação), R^2_α (coeficiente de determinação ajustado), além do Cp de Mallows dos modelos.

A seleção de variáveis foram feitas a partir de três modelos, o Modelo 1, exibido aqui, o Modelo 2, em que a variável MPG na cidade só foi ajustada até o grau 3, e o Modelo 3, em que ela foi ajustada até o grau 2. Vale dizer que mesmo para estes modelos, procurou-se evitar que variáveis muito correlacionadas estivessem ambas presentes, apesar de que, em alguns momentos, considerando que a seleção de variáveis seria feita, preferiu-se mantê-las.

Ao final do *Stepwise*, o ajuste selecionado considerava o a largura (grau dois), a localização do motor, MPG na cidade (também em grau dois), o tipo de motor, as rodas motrizes, o estilo de corpo do carro, o peso de meio fio, o número de cilindros e a taxa de compressão. Para estas variáveis, o coeficiente de determinação ajustado foi de 0,916, o Cp de Mallows foi 12,099, o AIC foi 3548,7146, a raiz quadrada do erro quadrático médio foi 2353,2733 e o BIC foi de 3623,637, o mais baixo dentre os modelos selecionados a partir do Modelo 3.

Vale dizer, que o fato deste ser o BIC mais baixo dentre estes modelos pesou na seleção das variáveis, tendo em vista que, apesar de como o AIC, o BIC ser um penalizador, ele penaliza mais que o primeiro. Assim sendo, a escolha pelas variáveis foi feita considerando os problemas de multicolinearidade vigentes e, portanto, se apoiando mais no critério que penalizava mais.

Após diversos testes ordenados, verificou-se que algumas das variáveis não significativas em termos preditivos, ajudavam com os pressupostos do modelo (que ainda estavam relativamente problemáticos), foram elas a distância entre eixos (*wheel_base*), o número de portas e a aspiração.

Somado a isso, precisou-se realizar uma transformação Box-Cox na variável preço; o parâmetro selecionado foi tal que $\lambda = -0,1818...$ e, assim, o modelo final ficou como a na Tabela 4.

Tabela 4 – Modelo Final

Parâmetro	Coefficiente	Erro Padrão	Valor - p
Intercepto	4,4085828	0.0773059	0,00
Largura	0,1994778	0.0672217	0,00
Localização do motor (traseira)	0.0910047	0.0230869	0,00
MPG na cidade	-0,6676684	0.0672669	0,00
MPG na cidade ²	0.2437921	0.0351364	0,00
Tipo de motor (1)	-0.0309861	0.0131761	0,01
Tipo de motor (ohc)	0.0099351	0.0090865	0,27
Tipo de motor (ohcf)	-0.0032406	0.0130196	0,80
Tipo de motor (ohcv)	-0.0178063	0.0133840	0,19
Rodas motrizes (fwd)	-0.0060400	0.0114646	0,60
Rodas motrizes (rwd)	0.0323662	0.0126998	0,01
Estilo de corpo (hardtop)	-0.0472742	0.0152140	0,00
Estilo de corpo (hatchback)	-0.0469811	0.0132184	0,00
Estilo de corpo (sedan)	-0.0394083	0.0138887	0,00
Estilo de corpo (wagon)	-0.0500949	0.0153346	0,00
Número de cilindros (1)	0.0187099	0.0095126	0,05
Número de cilindros (2)	0.0414069	0.0250468	0,1
Taxa de compressão (1)	0.0544959	0.0110418	0,00
Aspiração (turbo)	0.0049722	0.0068665	0,47
Distância entre eixos	0.0010634	0.0008073	0,19
Número de portas	-0.0098026	0.0064336	0,13
R ²	0,9263	R ² _α	0,9171
Estatística F	101,1 em 21 e 169 G.L	Valor-p:	≈ 0

Fonte: Autores (2021)

A interpretação do erro padrão, dos valores-p e da estatística F é como anteriormente. Vale salientar, no entanto, que como a transformação Box-Cox foi realizada na variável resposta, a variação média do preço em dólares não é mais feita diretamente.

A transformação Box-Cox é dada da seguinte maneira:

$$\frac{y_i^\lambda - 1}{\lambda}, \text{ se, } \lambda \neq 0;$$

$$\ln(y_i), \text{ c.c.}, \forall i \in 1, \dots, n;$$

Assim sendo, é possível obter a real alteração do preço substituindo λ pelo valor citado anteriormente e obtendo a função inversa da transformação Box-Cox; o que, apesar de aumentar o trabalho interpretativo, não é difícil de ser implementado e fornece benefícios em termos da análise de resíduos, como poderá ser visto a partir de agora.

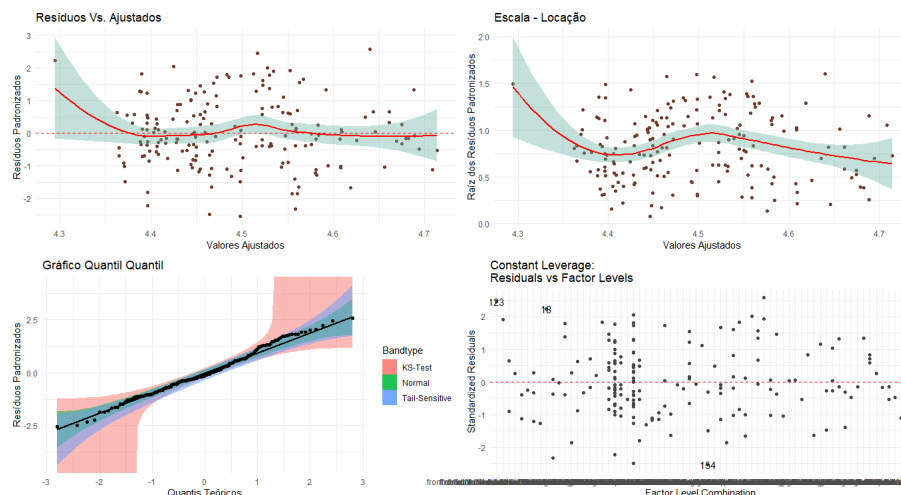


Figura 5 – Análise de resíduos do modelo final - Parte 1

Fonte: Autores (2021)

Como é possível ver na Figura 5, no gráfico de Resíduos Vs. Ajustados, apesar de uma heterocedasticidade não ser necessariamente observada, o primeiro ponto da esquerda para a direita chamou atenção. O comportamento foi similar no gráfico de escala-locação, enquanto que no gráfico de Leverage algumas das observações preocupantes são indicadas. Dito isso, o gráfico quantil quantil, no gráfico quantil quantil, os resíduos não parecem fugir do esperado de uma distribuição normal; o que é positivo.

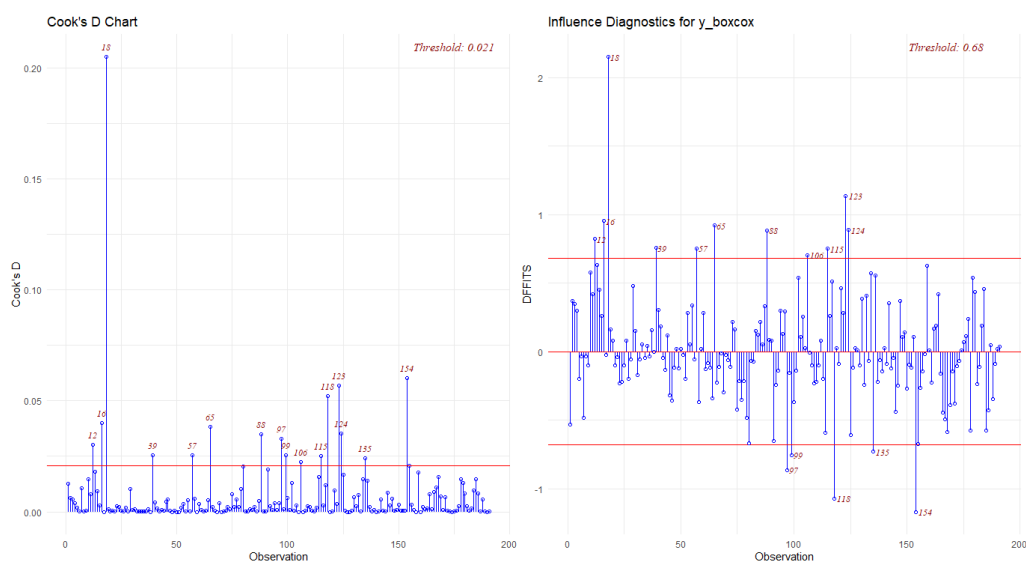


Figura 6 – Análise de resíduos do modelo final - Parte 2

Fonte: Autores (2021)

Na Figura 6, observam-se gráficos que medem a influência dos pontos na modelagem. É possível perceber que a observação identificada com o número 18 é, claramente, um ponto que parece não ter sido bem ajustado. O que é corroborado pela Figura 7, em que é possível observar que este, é tanto um *outlier* (um valor extremo), quanto um ponto de alavanca e, portanto, além de atípico é influente. Além disso, é possível observar que outros diversos pontos foram classificados como influentes, e alguns poucos como *outliers*.

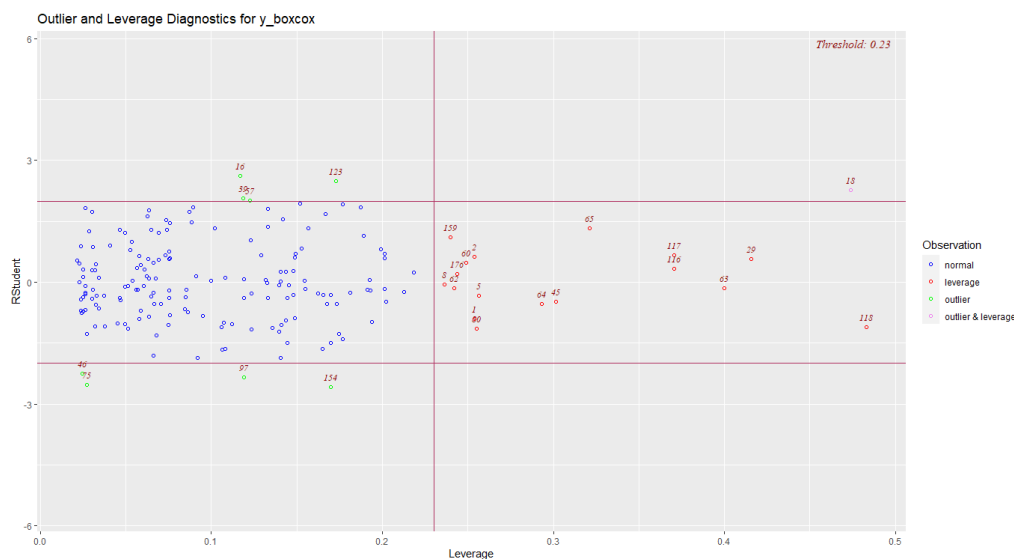


Figura 7 – Análise de resíduos do modelo final - Parte 3

Fonte: Autores (2021)

As outras medidas de influência visualizáveis poderão ser acessadas no código da análise de dados. Em geral, os DFFBetas corroboram o que foi mencionado até aqui; apesar de que no ajuste de algumas das variáveis, os pontos que fugiram do esperado foram os 159 e o 123.

Assim sendo, a seguir será possível visualizar os resultados dos testes de hipótese, como feito para o Modelo 1.

Tabela 5 – Análise de resíduos - Modelo final

Pressuposto	Teste	Estatística	Valor-p
Normalidade	Shapiro-Wilk	0,9891	0,09425
Homoscedasticidade	Breush-Pagan	26,04	0,2432
Independência	Durbin Watson	1,5149	0
Multicolinearidade	Média dos VIFs	2,811306	-

Fonte: Autores (2021)

Como é possível observar na Tabela 5, o novo modelo ajustado passa tanto nos testes de normalidade quanto de homoscedasticidade, ao nível de 5%. Além disso, a média dos VIFs agora é próxima de 3, o que por si só não é suficiente como indicativo; no entanto, ao olhar os VIFs e a tolerância de cada uma das variáveis ajustadas, pôde-se perceber que o maior dos VIFs era o da variável tipo de motor, com 8,547064. Pela regra prática, valores menores que 10 indicam a não presença de multicolinearidade e, portanto, no que concerne à esta medida, o modelo parece bem ajustado.

Quanto à normalidade, realizaram-se mais alguns testes; retirando a variável de localização do motor (que não é significativa, segundo o teste t), o valor-p do teste de hipóteses aumenta, no entanto, perde-se mais em termos da homoscedasticidade, do que se ganha com a normalidade. Outra possibilidade averiguada foi a retirada do ponto influente identificado pelo número 18. Apesar de realmente o ajuste em termos de normalidade melhorar, avaliou-se que a retirada deste ponto não melhorava o ajuste suficientemente para que o modelo final fosse ajustado sem ele. Além disso, como o ajuste passou pelo teste de normalidade nos níveis mais usuais (e, como a análise visual indica a normalidade dos resíduos), se decidiu não tentar melhorar mais o ajuste com este fim.

Finalmente, quanto ao teste de Durbin-Watson para independência, é possível notar que o valor-p segue

indicando que há uma autocorrelação entre os resíduos. Apesar disso, a regra prática indicam que valores de estatística de teste entre 1,5 e 2,5, no geral, são suficientes para que o modelo seja considerado independente. E, portanto, assim foi feito. Vale salientar, no entanto, que observou-se uma espécie de *trade-off* entre bons resultados na multicolinearidade e na independência.

Problemas com a independência do modelo podem ser resultantes de uma não especificação de variáveis importantes para explicar o preço, enquanto os problemas com multicolinearidade aumentam a medida que variáveis correlacionadas entre si são acrescentadas (e, nestes dados, essa é basicamente a regra para as variáveis bem relacionadas com o preço). Assim sendo, o acréscimo de algumas variáveis que melhoravam a independência, pioravam a multicolinearidade; e a solução para isso foi acrescentar algumas variáveis não tão correlacionadas com o preço (diretamente), mas que também não estavam tão correlacionadas com as outras variáveis e, além disso, ajudavam a diminuir os problemas com independência do ajuste. Na Tabela 4, é fácil ver que estas, são as últimas variáveis adicionadas.

4 CONCLUSÕES

Levando em conta tudo o que foi considerado, é importante lembrar que a análise foi feita com intuito de encontrar ajustar um modelo de regressão linear múltiplo, entre as variáveis preditoras e o preço na base de dados *Import-85*, disponibilizada por Dua e Graff (2017). A análise de dados procurou levantar hipóteses sobre a importância das variáveis, obter bons resultados em termos dos pressupostos do modelo e, inclusive, precisou ser corrigida diversas vezes para lidar com a normalidade, homoscedasticidade, independência e multicolinearidade.

O modelo final foi ajustado após um procedimento de seleção de variáveis via *Stepwise*, em que se avaliaram C_p de Mallows, R^2 , R_a^2 , erro quadrático médio, AIC e BIC. Fez-se necessário retirar algumas variáveis e acrescentar outras para obtenção de melhores resultados em termos de pressupostos.

Ainda assim, melhorias em diversos pontos podem ser feitas. Para que não se tenha tantas variáveis, - aparentemente informativas - não utilizadas no modelo final, abordagens mais complexas para lidar com multicolinearidade poderiam ter sido empregadas. Além disso, os problemas com a independência poderiam ter sido corrigidos, por exemplo, através de modelos auto-regressivos.

Uma outra possibilidade seria tentar abordar os dados a partir de uma perspectiva multivariada, ou através de um outro modelo que não o linear. Dentro das possibilidades do modelo linear, algumas das variáveis com categorias em que apenas um indivíduo se fazia presente poderiam ter sido recategorizadas, para que a modelagem destes fosse facilitada. Um aumento no tamanho da amostra dessas categorias mais raras também seria proveitoso, pois permitiria o acréscimo de termos de interação, que possivelmente levariam a uma modelagem mais precisa.

Finalizando, em geral as suposições do modelo foram cuidadosamente verificadas, com ressalvas para a independência dos resíduos. O modelo foi considerado significativo e seu coeficiente de determinação ajustado foi calculado acima de 0,9; de modo que o ajuste foi considerado satisfatório.

REFERÊNCIAS

- BÉCUE-BERTAUT, M.; PAGÈS, J. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. **Computational Statistics & Data Analysis**, Elsevier, v. 52, n. 6, p. 3255–3268, 2008.
- DEMÉTRIO, C. G. B.; ZOCCHI, S. S. Modelos de regressão. **Piracicaba: ESALQ**, 2006.
- DING, X. et al. Using structured events to predict stock price movement: An empirical investigation. In: **PROCEEDINGS of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. P. 1415–1425.
- DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. [S.l.: s.n.], 2017. Disponível em: [🔗](#).
- FERREIRA, K. A.; RIBEIRO, P. C. C. Tecnologia da informação e logística: os impactos do EDI nas operações logísticas de uma empresa do setor automobilístico. **XXIII ENEGEP-Encontro Nacional de Engenharia de Produção, Ouro Preto**, 2003.
- GONZALEZ, R. V. D.; MARTINS, M. F. Melhoria contínua e aprendizagem organizacional: múltiplos casos em empresas do setor automobilístico. **Gestão & Produção**, SciELO Brasil, v. 18, p. 473–486, 2011.
- _____. Melhoria contínua no ambiente ISO 9001: 2000: estudo de caso em duas empresas do setor automobilístico. **Production**, SciELO Brasil, v. 17, p. 592–603, 2007.
- HOFFMANN, R.; VIEIRA, S. Análise de regressão: uma introdução à econometria. **São Paulo**, 1998.
- JAMES, G. et al. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112.
- RAWLINGS, J. O.; PANTULA, S. G.; DICKEY, D. A. **Applied regression analysis: a research tool**. [S.l.]: Springer Science & Business Media, 2001.
- RESENDE, M. et al. Melhor predição linear não viciada (Blup) de valores genéticos no melhoramento de Pinus. **Embrapa Florestas-Artigo em periódico indexado (ALICE)**, Boletim de Pesquisa Florestal, Colombo, n. 32/33, p. 3-22, jan./dez. 1996., 1996.
- VITTINGHOFF, E. et al. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. Springer, 2006.

ANEXO A – ACESSO A ANÁLISE DE DADOS

(Incluindo acesso a base de dados, *scripts* e códigos de análise, gráficos demonstrados aqui e códigos para plotagem dos outros gráficos. Além de alguns outros gráficos analisados)

https://github.com/Mkyou/labs-regressao/tree/main/lab_final

ANEXO B – ENDEREÇOS QUE FORAM INTERESSANTES PARA O TRABALHO

<https://online.stat.psu.edu/stat501/lesson/14/14.1>

<https://www.r-graph-gallery.com/ggplot2-package.html>

<https://www.rdocumentation.org/packages/olsrr/versions/0.5.3>

AGRADECIMENTOS

O(s) autor(es) agradece(m) à Universidade Federal da Bahia pelo apoio recebido para o desenvolvimento do presente trabalho.

DECLARAÇÃO DE RESPONSABILIDADE

O(s) autor(es) é(são) o(s) único(s) responsável(eis) pelas informações contidas neste documento.