Ministério da Educação
Universidade Federal da Bahia
Campus Ondina
Departamento de Estatística - UFBA

LABORATÓRIO FINAL - VERSÃO 1.1
FINAL LABORATORY - VERSION 1.1

Gabriel Penha*, Moisés Augusto†, Marla Lorrani‡

RESUMO

A base de dados *import-85*, obtida no repositório de dados para aprendizado de máquina da Universidade da Califórnia, continha 205 observações sobre 26 variáveis de interesses de diversos modelos de carros. Este trabalho objetiva analisar tais dados, com intuito de verificar as relações das diversas características dos carros, com seu respectivo preço - a variável resposta -. Por ora, o tratamento de dados e uma breve análise exploratória foram feitas, além de ajustes de alguns modelos lineares simples, que serão descritos e de um modelo múltiplo, através do método de seleção de ajuste *backward*, que apesar de não ser descrito, apresentou algumas ressalvas que foram elencadas neste documento.

Palavras-chave: Carros. Preço. Modelos lineares. Tratamento. Transformações.

1 INTRODUÇÃO

Na base de dados *import-85*, disponibilizada pelo Dua e Graff (2017), encontra-se informações sobre diversas características de 205 diferentes carros. Entre elas, a fabricante do carro, o tipo de combustível e o número de portas podem ser citados como exemplo.

Ao todo, são 26 variáveis, incluindo a variável resposta preço, que indica o preço do carro em Unidades Monetárias ($U.M$). Algumas delas com um número considerável de informações faltantes, outras com nomes-base não muito adequados para análises de dados em computadores e ainda, todas elas com classe de variáveis incorreto; isto é, todas as variáveis eram do tipo *String*, incluindo as numéricas, o que também não é tão indicado para este tipo de análise.

Dito isso, antes que uma análise exploratória fosse iniciada, fez-se necessário realizar um tratamento de dados, explicitado na curta seção seguinte. Após ela, o leitor irá se deparar com uma seção dedicada à uma análise exploratória primária dos dados e, em seguida, a alguns modelos lineares ajustados entre as variáveis linear e significativamente correlacionadas com a variável resposta. Apesar de um ajuste de modelo múltiplo já ter sido realizado através do critério de seleção *backward*, ele não será descrito nesta versão. Finalmente, a última seção se dedicará à considerações finais.

2 TRATAMENTO DE DADOS

Primeiramente, é importante salientar que a base de dados *import-85* estava subdividida em dois arquivos *.Data*. O primeiro deles continha os dados, o segundo continha o nome das variáveis e algumas informações sobre elas. Deste modo, inicialmente foi necessário converter os arquivos *.Data* para arquivos *.csv*; posteriormente, criou-se um vetor de cadeias de texto - que serviriam como nome das variáveis - para que a base que continha, de fato, os dados, pudesse ser nomeada.

A maioria dos nomes utilizados foram os mesmos disponíveis no segundo arquivo *.Data* mencionado; com apenas algumas modificações tangentes à boas práticas de código, até mesmo para facilitar a análise de dados posterior.

* Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelando em Estatística; penha.gabriel@ufba.br.
† Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelando em Estatística; moises.augusto@ufba.br.
‡ Instituto de Matemática e Estatística, Departamento de Estatística, Bacharelando em Estatística; marla.lorrani@ufba.br.

Feito isso, a segunda coluna de observações, intitulada *normalized_losses* foi descartada, pois continha muitas observações faltantes. Além disso, fez-se necessário converter as colunas numéricas para *double* e as categóricas para *factor*. Também foi necessário transformar os valores faltantes da base de dados, que estavam na forma "?" em *NAs* clássicos, entendidos pelo *software* estatístico R. Após isso, tais observações foram removidas.

Ao final do processo, a base de dados tratada possuía 193 linhas e 25 colunas. E com ela, iniciou-se a análise exploratória dos dados - preeliminar -.

3 ANÁLISE EXPLORATÓRIA

Como mencionado anteriormente, o conjunto de dados final possuía 25 colunas e 193 linhas; por representarem informações bastante diferentes, as variáveis encontravam-se em escalas diferentes. Por questões de espaço e praticidade, este relatório não as descreverá por inteiro - somente uma ou outra que forem destacadas -; no entanto, no endereço-*web* disponibilizado nos Anexos deste documento, o projeto da análise destes dados estará disponível, bem como um arquivo com a descrição de cada uma das co-variáveis (de nome *import-85-names.csv*.

Dito isto, toda a análise exploratória visual e individual também estará lá disponível; aqui, apenas ressaltaremos as informações mais importantes.

Na Figura 1, é possível visualizar as estatísticas descritivas para cada co-variável presente na base de dados.

```
  symboling               make     fuel_type   aspiration  num_doors        body_style
Min.   :-2.0000   toyota    :32   diesel: 19   std  :158   four:112   convertible: 6
1st Qu.: 0.0000   nissan    :18   gas   :174   turbo: 35   two : 81   hardtop    : 8
Median : 1.0000   honda     :13                                       hatchback  :63
Mean   : 0.7979   mitsubishi:13                                       sedan      :92
3rd Qu.: 2.0000   mazda     :12                                       wagon      :24
Max.   : 3.0000   subaru    :12
                  (Other)   :93
drive_wheels engine_location   wheel_base         length          width
4wd:  8      front:190       Min.   : 86.60   Min.   :141.1   Min.   :60.30
fwd:114      rear :  3       1st Qu.: 94.50   1st Qu.:166.3   1st Qu.:64.10
rwd: 71                      Median : 97.00   Median :173.2   Median :65.40
                             Mean   : 98.92   Mean   :174.3   Mean   :65.89
                             3rd Qu.:102.40   3rd Qu.:184.6   3rd Qu.:66.90
                             Max.   :120.90   Max.   :208.1   Max.   :72.00

    height       curb_weight   engine_type num_cylinders  engine_size     fuel_system
Min.   :47.80   Min.   :1488   dohc : 12   eight :  4    Min.   : 61.0   mpfi   :88
1st Qu.:52.00   1st Qu.:2145   dohcv:  0   five  : 10    1st Qu.: 98.0   2bbl   :64
Median :54.10   Median :2414   l    : 12   four  :153    Median :120.0   idi    :19
Mean   :53.87   Mean   :2562   ohc  :141   six   : 24    Mean   :128.1   1bbl   :11
3rd Qu.:55.70   3rd Qu.:2952   ohcf : 15   three :  1    3rd Qu.:146.0   spdi   : 9
Max.   :59.80   Max.   :4066   ohcv : 13   twelve:  1    Max.   :326.0   mfi    : 1
                               rotor:  0   two   :  0                   (Other): 1
     bore           stroke      compression_ratio   horsepower       peak_rpm
Min.   :2.540   Min.   :2.070   Min.   : 7.00     Min.   : 48.0   Min.   :4150
1st Qu.:3.150   1st Qu.:3.110   1st Qu.: 8.50     1st Qu.: 70.0   1st Qu.:4800
Median :3.310   Median :3.290   Median : 9.00     Median : 95.0   Median :5100
Mean   :3.331   Mean   :3.249   Mean   :10.14     Mean   :103.5   Mean   :5100
3rd Qu.:3.590   3rd Qu.:3.410   3rd Qu.: 9.40     3rd Qu.:116.0   3rd Qu.:5500
Max.   :3.940   Max.   :4.170   Max.   :23.00     Max.   :262.0   Max.   :6600

   city_mpg      highway_mpg        price
Min.   :13.00   Min.   :16.00   Min.   : 5118
1st Qu.:19.00   1st Qu.:25.00   1st Qu.: 7738
Median :25.00   Median :30.00   Median :10245
Mean   :25.33   Mean   :30.79   Mean   :13285
3rd Qu.:30.00   3rd Qu.:34.00   3rd Qu.:16515
Max.   :49.00   Max.   :54.00   Max.   :45400
```

Figura 1 – Estatísticas descritivas: *Import-85*

Em que *Min.*, *1st Qu.*, *Median*, *Mean*, *3rd Qu.* e *Max.* indicam, respectivamente, o mínimo, o primeiro

quartil, a mediana, a média, o terceiro quartil e o máximo das observações para sua respectiva coluna (quando as variáveis são numéricas). É possível reparar que nenhuma destas colunas parecem possuir estatísticas muito fora do esperado. Além disso, não parecem haver indicativos muito fortes de assimetria em nenhuma das variáveis numéricas.

Nas colunas presentes na Figura 1 que não possuem as descritivas acima mencionadas, é possível ver a quantidade de cada fator para a respectiva variável; por exemplo, em *fuel_type*, 174 dos carros são movidos a gasolina, enquanto 19 deles são movidos a diesel.

Considerando tais informações, partiu-se para uma análise visual bivariada das características dos carros (considerando a variável resposta e uma a uma, cada uma das variáveis explicativas).

3.1 Análise visual

Após a análise da distribuição das variáveis e de suas descritivas, verificaram-se, visualmente, a relação bivariada entre o preço em unidades monetárias e cada uma das variáveis explicativas.

Na Figura 2, é possível visualizar o diagrama de dispersão entre a a característica nomeada por *engine_size* e o preço. É possível perceber fortes indícios de uma relação linear positiva entre as duas variáveis; isto é, parece que conforme *engine_size* cresce, o preço também cresce.

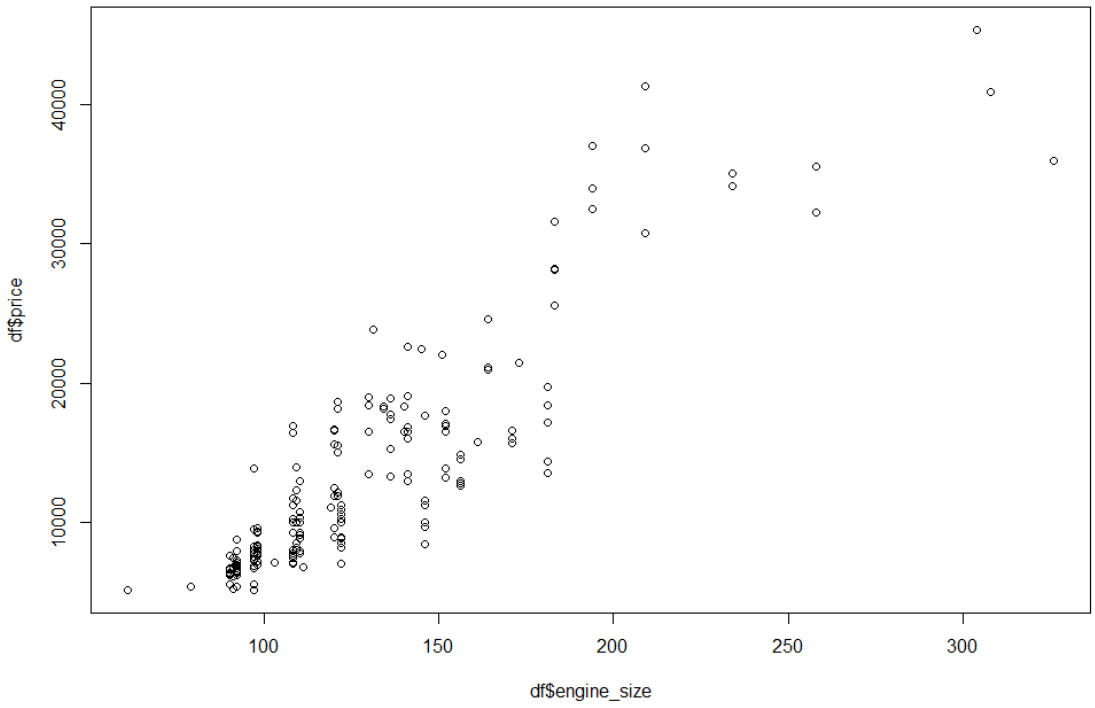

Figura 2 – *Engine size* Vs. Preço ($U.M$)

Dentre as características numéricas, a relação do preço com as seguintes variáveis: *wheel_base, height, bore, stroke, compression_ratio* e *peak_rpm* não parecia ser suficientemente significante, do ponto de vista linear; enquanto *length, width, curb_weight, engine_size, horsepower, city_mpg* e *highway_mpg* pareciam ser significantes.

Já considerando as co-variáveis categóricas, *num_doors* não parecia possuir uma relação muito significante

para com o preço, diferentemente de todas as outras.

Na Figura 3, é possível visualizar um *boxplot* que aponta para uma forte relação entre a variável *engine_location* e o preço.

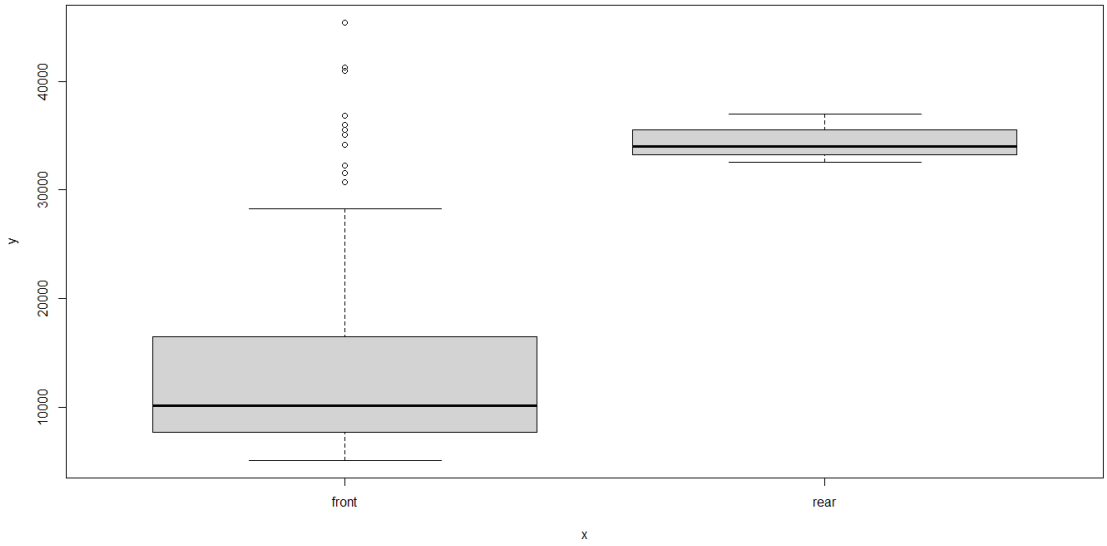

Figura 3 – *Engine location* Vs. Preço ($U.M$)

Como é possível perceber, os preços parecem ser mais altos quando a *engine* (o motor) é localizado na parte traseira do carro (*rear*).

4 RESULTADOS

4.1 Modelos lineares simples

Dentre as variáveis explicativas quantitativas, nem todas tiveram sua relação consideradas significativas - do ponto de vista linear - para com a variável resposta.

Para certo aprofundamento na análise, ajustaram-se alguns modelos lineares simples, considerando tais variáveis com relação linear significativa com o preço como explicativas, e, claro, o próprio preço, como variável resposta.

O Modelo 1 foi da seguinte forma:

$$\hat{Y} = -65358{,}17 + 451{,}13 length \tag{1}$$

Em que $\hat{Y}$ é o valor predito para o preço quando o comprimento do carro é igual a *length*. A cada aumento em unidades do comprimento, o preço médio sobe cerca de \$451,13. Salienta-se que este modelo só é válido para valores razoáveis da variável *length*; mais especificamente, o carro com menor comprimento considerado tinha 141,1 U.C (unidades de comprimento).

O Modelo 2 foi ajustado assim:

$$\hat{Y} = -174872{,}9 + 2855{,}5 width \tag{2}$$

Em que $\hat{Y}$ é como anteriormente e *width* é a largura do carro, também em U.C. A cada aumento em unidades do comprimento, o preço médio sobe cerca de \$2855,5. Salienta-se que este modelo só é válido para valores razoáveis da variável *width*; mais especificamente, o carro com menor largura considerado tinha 60,3 U.C (unidades de comprimento).

O Modelo 3 foi da seguinte forma:

$$\hat{Y} = -19580{,}00 + 12{,}83 curbWeight \tag{3}$$

Em que *curbWeight* é o *peso de meio fio* do carro e $\hat{Y}$ é como antes. A interpretação do modelo se dá como nos casos anteriores. Além disso, o modelo só é interpretável para valores razoáveis de peso.

A seguir, o Modelo 4:

$$\hat{Y} = -8862{,}79 + 172{,}86 engineSize \tag{4}$$

Em que *engineSize* é o tamanho do motor do carro e $\hat{Y}$ é como antes. A interpretação do modelo se dá como nos casos anteriores. Além disso, o modelo só é interpretável para valores razoáveis de tamanho do motor.

Modelo 5:

$$\hat{Y} = -4630{,}70 + 173{,}13 horsepower \tag{5}$$

Modelo 6:

$$\hat{Y} = 35947{,}36 + -894{,}81 cityMpg \tag{6}$$

E, finalmente, o Modelo 7:

$$\hat{Y} = 39558{,}85 + -853{,}39 highwayMpg \tag{7}$$

Em que, $\hat{Y}$ é como antes para todos os modelos, *horsepower* é o número de cavalos de força, *cityMpg* é a quantidade de milhas por galão que o carro faz na cidade e *highwayMpg* é o mesmo, porém em rodovias ao invés de cidades.

Todos os 7 modelos foram considerados significativos ao nível de 95%, e só são válidos para valores razoáveis de suas respectivas variáveis explicativas.

Além disso, os Modelos 1, 2, 3, 5, 6 e 7 não apresentaram bons resultados nas análises de resíduo (tendências fortes de não normalidade e heterocedasticidade), além de alguns pontos de alavanca possivelmente influentes, enquanto o Modelo 4, apesar de homocedástico, fugia da normalidade.

Estes resultados indicam que, possivelmente, parte da variabilidade do preço não está sendo capturada pelos ajustes considerados e, portanto, um modelo linear múltiplo poderia ser viável.

4.2 Modelos lineares múltiplos

Considerando as variáveis moderada (ou forte) e linearmente correlacionadas ao preço, além das variáveis categóricas com diferenças significativas de preço para com suas categorias, ajustou-se um modelo linear múltiplo através do critério de seleção de ajustes (*backward*), que, em linhas gerais, parte de um modelo cheio e vai retirando variáveis, uma a uma, para verificar qual dos possíveis ajustes seria o melhor.

No entanto, na variável explicativa categórica que denota o número de cilindros, um dos fatores - especificamente àquele que indica que o carro possui três cilindros - tem somente uma observação, o que comprometeu o ajuste do modelo considerando esta característica - observações *NA* foram produzidas -. Dito isso, as análises dos pressupostos posteriores foram prejudicadas.

Um segundo modelo foi ajustado, retirando tal variável; este apresentou não normalidade e heterocedasticidade, o que pode indicar que ainda existem características a se considerar (por exemplo, variáveis que possuem relação polinomial com o preço); a identificação dessas características e o ajuste de novos modelos são os próximos passos.

5 CONSIDERAÇÕES FINAIS

Considerando o mencionado, vale dizer que a análise exploratória ainda será refinada com testes de hipóteses que evidenciarão - ou não - o que foi sugerido pela análise visual; com gráficos um pouco mais apresentáveis e, talvez, com mais algumas análises que forneçam indícios de relacionamento entre as variáveis (que não necessariamente o preço).

Como mencionado, nenhum dos modelos simples ajustados apresentou bons resultados em termos de homocedasticidade e normalidade; além disso, somente os modelos que consideraram o tamanho do motor (*engine_size*) e a variável *curb_weight* explicavam mais que 70% da variabilidade do preço. Indícios que apontam para a necessidade de ajustes de modelos que incluam mais variáveis.

Além disso, o ajuste múltiplo desconsiderando relações possivelmente polinomiais também não performou bem diante aos pressupostos; o que fortalece a ideia referente a refinação da análise exploratória (a fim de identificar tais relações) e ajustes de novos modelos.

REFERÊNCIAS

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. [S.l.: s.n.], 2017. Disponível em: .

ANEXO A – CÓDIGOS UTILIZADOS PARA ANÁLISE (NO R)

É possível obter o projeto utilizado para análise de dados no R com o *link* a seguir:
https://github.com/Mkyou/labs-regressao

DECLARAÇÃO DE RESPONSABILIDADE

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