ANÁLISE DE REGRESSÃO PARA CARROS USADOS EM REVENDA

Aluno: Lucas Manoel Batista de Albuquerque*
Professor: Cleanderson Romualdo Fidelis*

Campina Grande – PB

2022

* Lucas Manoel Batista de Albuquerque

RESUMO

Trata-se de um estudo onde foi feito um modelo de regressão linear para tentar estimar o preço atual de um carro usado de acordo com seu tipo de combustível, preço que foi vendido pela primeira vez e quilometragem. O modelo que mais se adequou aos dados foi aquele que usou apenas das variáveis de quilometragem percorrida e o preço que o carro foi vendido a primeira vez, tendo um $R^2 = 0.803$, mas com os resíduos não normais. Foi feita posteriormente uma análise descritiva para identificar os carros que causaram essa não normalidade dos resíduos. O modelo em questão é descrito como $Y = -761{,}3211 + 1{,}4860 * (PrecoVenda) + 0{,}0396 * (KmDirigidos)$.

Palavras-chave: regressão linear; modelos lineares; carros; descritiva.

1 METODOLOGIA

1.1 SOBRE OS DADOS

Contém informações sobre carros usados, no qual tentaremos prever seus preços atuais através de suas covariáveis usando regressão linear. As covariáveis são: nome do veículo (não será usada para estimar um modelo), preço que foi vendido em reais, preço atual em reais (variável resposta), quilômetros dirigidos, tipo de combustível (gasolina ou diesel).

1.2 REGRESSÃO LINEAR POR MÍNIMOS QUADRADOS

A análise de regressão linear é usada para prever o valor de uma variável com base no valor de outra. A variável que deseja prever é chamada de variável dependente. A variável que é usada para prever o valor de outra variável é chamada de variável independente.

A regressão linear é dada pela equação 1:

$$\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{e} \tag{1}$$

Onde $\boldsymbol{Y}$ é a variável dependente, $\boldsymbol{X}$ a matriz de constantes conhecidas ou variáveis independentes, $\boldsymbol{\theta}$ vetor de elementos desconhecidos que desejamos estimar e $\boldsymbol{e}$ o erro aleatório associado que assumiremos que segue uma distribuição normal.

Escrevendo esse mesmo modelo para que vejamos os elementos de cada matriz, tal qual na equação 2, temos:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{2}$$

Onde a primeira matriz é o vetor de observações, a segunda são as constantes conhecidas, a terceira são os parâmetros a serem estimados e a última o vetor de erros associados.

Para obtermos os valores da matriz de parâmetros estimados, basta isolarmos a matriz $\boldsymbol{\theta}$, ou a matriz de betas, na equação matricial. Sendo assim, obtemos o seguinte resultado:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-}(\boldsymbol{X}^T\boldsymbol{Y}) \tag{3}$$

No qual a primeira matriz, denominada por $\widehat{\boldsymbol{\theta}}$, é a matriz de parâmetros estimados, $\boldsymbol{X}^T$ é a matriz de constantes conhecidas transposta, $\boldsymbol{X}$ é a matriz de

constantes conhecidas, $-$ é a função para encontrar a matriz inversa e $\boldsymbol{Y}$ a matriz de valores observados. Temos então a matriz dos parâmetros estimados para cada covariável.

1.3 CRITÉRIO DE ESCOLHA DAS VARIÁVEIS INDEPENDENTES

Iremos testar a hipótese nula de um parâmetro estimado β_k ser igual a zero, contra a hipótese desse mesmo parâmetro ser diferente de zero. Considerando isto, temos, sob a hipótese nula:

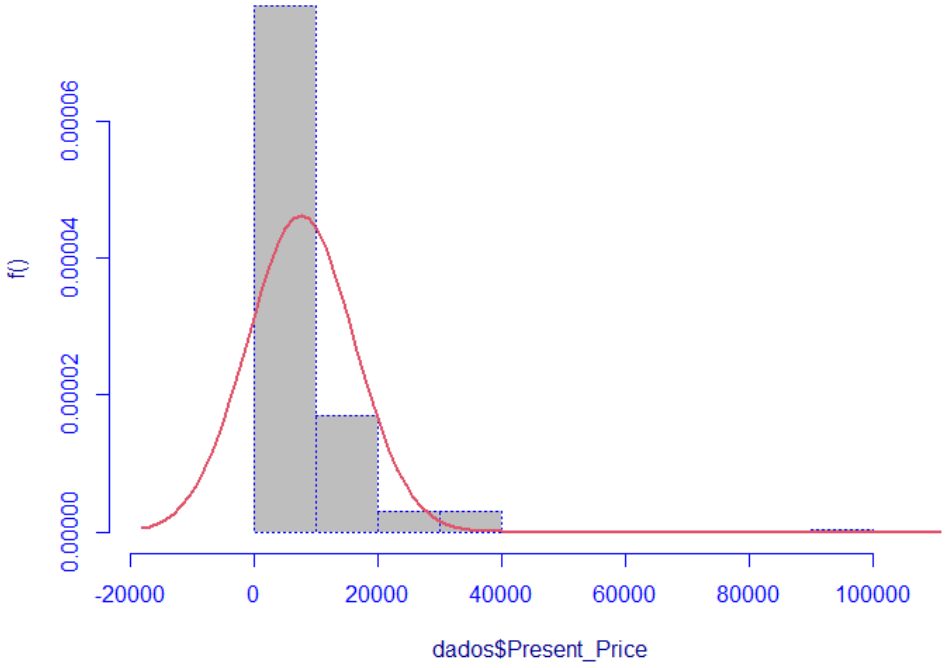
$$\frac{\widehat{\beta_k}}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim t(n-2) \tag{4}$$

Sendo $\widehat{\sigma^2}$ a estimativa variância do erro, descrita por $\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i-\widehat{y_i})^2}{n-2}$. Assim, para um teste de nível 5% rejeitamos a hipótese nula, se o módulo da estatística em (4) for maior do que $t(\frac{5\%}{2}, n-2)$.

2 RESULTADOS E DISCUSSÃO

Após a apresentação dos dados e a apresentação da metodologia utilizada, partiremos para uma breve análise descritiva dos dados. A seguir, na Figura 1, a distribuição do preço da variável dependente “preço atual em reais”:

Figura 1 - Distribuição do preço atual dos carros usados



Fonte: O Autor.

Abaixo, na Tabela 1 e Tabela 2, algumas estatísticas descritivas dos dados que serão utilizados para realizar a regressão linear:

Tabela 1 – Estatísticas descritivas das variáveis usadas para estimação do modelo de regressão

Variável	Mínimo	Máximo	Média	Mediana
Preço venda	100	35000	4661.296	3600
preço atual	320	92600	7628.472	6400
km dirigidos	500	500000	36947.206	32000

Fonte: O autor.

Tabela 2 – Estatísticas descritivas das variáveis usadas para estimação do modelo agrupadas por tipo de combustível

Combustível	Variável	Mínimo	Máximo	Média	Mediana
Diesel	Preço venda	2950	35000	10046.935	7600
Diesel	preço atual	5090	92600	15511.290	10380
Diesel	km dirigidos	2071	197176	50124.081	45000
Petrol	Preço venda	100	19750	3264.184	2650
Petrol	preço atual	320	23730	5583.556	4600
Petrol	km dirigidos	500	500000	33528.937	25870

Fonte: O autor.

Seguindo com a análise, faremos a aplicação do modelo de regressão para estimar os parâmetros de cada modelo, assim como a sua significância associada. Temos então que as variáveis significativas foram o preço em que o carro foi

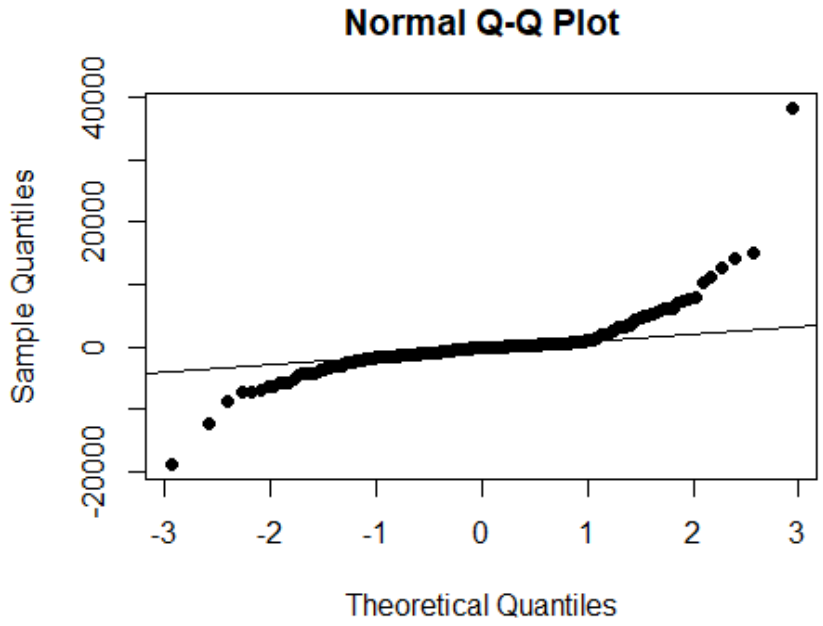
vendido e a sua quilometragem. O modelo estimado está descrito de acordo com a equação (5):

$$Y = -761{,}3211 + 1{,}4860 * (PrecoVenda) + 0{,}0396 * (quilometragem) \quad (5)$$

Temos um modelo de comportamento positivo, ou seja, na medida em que o preço em que o carro foi vendido no passado e a sua quilometragem aumentam, o preço atual do carro também tende a ser maior.

Desse mesmo modelo estimado podemos averiguar o “qqplot” dos seus resíduos, afim de identificar possíveis variáveis que destoam dos valores que o modelo previu.

Figura 1 - Gráfico qqplot dos resíduos do modelo estimado



Fonte:O autor.

Selecionando os carros que tiveram um erro maior do que 10000 na previsão do modelo, temos:

Tabela 3 – Veículos com erro maior do que 10000 reais na previsão do preço atual

Mome do Carro	Quilometragem	Preço venda	Preço atual	Valores preditos
corolla altis	50000	4750	18540	8277.169
fortuner	6000	33000	36230	48514.428
corolla altis	80000	5250	22830	10208.152
innova	15000	23000	25390	34010.776
camry	142000	2500	23730	8576.801
land cruiser	78000	35000	92600	54337.591
corolla altis	62000	3800	18610	7340.657
corolla altis	89000	4000	22780	8707.041
Activa 3g	500000	170	520	19290.983

Da Tabela 3, vemos que os carros corolla altis, innova, land cruiser e camry valorizaram bastante desde a última vez que foram vendidos, fazendo com que o modelo não reconheça tal comportamento. O carro Activa 3g teve um preço predito muito acima do que ele está agora, pois sua quilometragem é a mais alta que podemos encontrar entre todos os carros e está muito acima da média, e como a quilometragem tem uma influência positiva sobre o resultado do modelo, obtivemos um resultado bem distante da realidade.

3 CONSIDERAÇÕES FINAIS

Observamos que existe uma relação positiva entre os preços atuais dos carros usados e os preços em que eles foram vendidos no passado junto com sua quilometragem. O modelo de regressão se ajustou bem aos dados, salvo algumas exceções de carros que se valorizaram muito ao longo do tempo ou que têm uma quilometragem muito acima da média.

REFERÊNCIAS

CHARNET, Reinaldo; FREIRE, C.A. de L.; CHARNET, E.M.R; BONVINO, Heloísa; *Análise de modelos de regressão linear: Com aplicações*. Segunda edição. Universidade Estadual de Campinas: Editora Unicamp, maio de 2015.

R Core Team. R: *A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>.

ANEXO A

```
library(tidyverse)
dados <- read.csv("car data.csv")
dados$Selling_Price <- 1000*dados$Selling_Price
dados$Present_Price <- 1000*dados$Present_Price
require(dplyr)

        estatisticas = dados %>% summarise(VARIAVEL = c("Preco venda",'preco atual','km dirigidos'),
                                           minimo = round(c(min(dados$Selling_Price),min(dados$Present_Price),min(dados$Kms_Driven)),3),
                                           maximo = round(c(max(dados$Selling_Price),max(dados$Present_Price),max(dados$Kms_Driven)),3),
                                           media = round(c(mean(dados$Selling_Price),mean(dados$Present_Price),mean(dados$Kms_Driven)),3),
```

```
                                mediana                                 =
round(c(median(dados$Selling_Price),median(dados$Present_Price),median(dados$
Kms_Driven)),3))

        knitr::kable(estatisticas,"pipe",align = "c",
          caption = "Descritivas")

        estatisticas2                   =                   dados                   %>%
dplyr::select(Selling_Price,Present_Price,Kms_Driven,Fuel_Type) %>%
          group_by(Fuel_Type) %>% summarise(VARIAVEL = c("Preco venda",'preco
atual','km dirigidos'),
                                minimo                                  =
round(c(min(Selling_Price),min(Present_Price),min(Kms_Driven)),3),
                                maximo                                  =
round(c(max(Selling_Price),max(Present_Price),max(Kms_Driven)),3),
                                media                                   =
round(c(mean(Selling_Price),mean(Present_Price),mean(Kms_Driven)),3),
                                mediana                                 =
round(c(median(Selling_Price),median(Present_Price),median(Kms_Driven)),3))

kableExtra::kable(estatisticas2,"pipe", align="c", caption = "Descritivas por tipo de
combustível")

        petrolio <- ifelse(dados$Fuel_Type=="Petrol",1,0)
        diesel <- ifelse(dados$Fuel_Type=="Diesel",1,0)
        manual <- ifelse(dados$Transmission =="Manual",1,0)
        automatica <- ifelse(dados$Transmission=="Automatic",1,0)
        individual <- ifelse(dados$Seller_Type=="Individual",1,0)
        distribuidora <- ifelse(dados$Seller_Type=="Dealer",1,0)
        df_reg <- data.frame(preco_venda = dados$Selling_Price,
                        preco_atual = dados$Present_Price,
                        diesel = diesel,
                        petrolio = petrolio,
                        Kms = dados$Kms_Driven,
                        manual = manual,
                        automatica = automatica,
                        individual = individual,
                        distribuidora=distribuidora)

        constante <- rep(1,times=301)
        x <- cbind(constante = constante,
                    preco_venda = dados$Selling_Price,
                    petrolio = petrolio,
                manual,
                    Kms = dados$Kms_Driven,
                individual = individual)

y<- dados$Present_Price

solve(t(x)%*%x)%*%(t(x)%*%y)
```

```
	model <-lm(dados$Present_Price ~ dados$Selling_Price+ dados$Kms_Driven)
	#summary(model)
	setwd("C:/Users/Lucas/Documents/modelos lineares/dataset")
	dados <- read.csv("car data.csv")
	dados$Selling_Price <- 1000*dados$Selling_Price
	dados$Present_Price <- 1000*dados$Present_Price
	#dados <- dados[-c(197),]
	#View(dados)
	model <-lm(dados$Present_Price ~ dados$Selling_Price+ dados$Kms_Driven)
	#summary(model)
	#plot(model)
	#fe <- qqnorm(model$residuals, pch=16)
	#qqline(model$residuals)
	#shapiro.test(model$residuals)
	#identify(fe)
	out_resd <- dados[c(58,65,79,83,86,87,91,95,197),c(1,5,3,4)]
	#out_resd
	#plot(model)
	require(gamlss)
gamlss::histDist(dados$Present_Price, main="Distribuição do preço atual dos carros usados")

	ident <- c(58,65,79,83,86,87,91,95,197)
	fe<-qqnorm(model$residuals,pch=16)
qqline(model$residuals)

	out_resd <-
cbind(out_resd,model$fitted.values[c(58,65,79,83,86,87,91,95,197)],mean(dados$Selling_Price),mean(dados$Kms_Driven))
	out_resd <- as_tibble(out_resd)

kableExtra::kable(out_resd,"pipe", align="c", caption = "", col.names = c("nome do carro", "kms dirigidos", "preco venda", "preco atual","valores preditos", "media preco de venda", "media km dirigidos"))
```