Universidade Estadual de Campinas

Instituto de Matemática, Estatística e Computação Científica

Projeto II de Análise de Regressão

19 de Outubro de 2021

Alex Dos Santos Lima - RA: 230560

ANÁLISE DE REGRESSÃO LINEAR MÚLTIPLA

Trabalho apresentado ao Curso de Estatística da
Universidade Estadual de Campinas (UNICAMP) como
requisito parcial para aprovação na disciplina de
ME613 de Análise De Regressão.

Universidade Estadual de Campinas

Professor Aluísio de Souza Pinheiro

Campinas

2021

1. Introdução ao conjunto de dados

O objetivo deste relatório é realizar uma análise de *Regressão Linear Múltipla* do pacote de dados disponível do Shoftware *R* cuja o nome da biblioteca é *Ecdat* sendo o pacote com o nome *Computers*, com 9 variável e 6259 observações do conjunto de dados. As variáveis que será explorada ao longo do relatório segue da seguinte forma:

- *Preço*: Preço em dólares americanos de 486 PCs
- *Velocidade*: Velocidade do clock em MHz
- *HD*: Tamanho do disco rígido em MB tamanho da memória RAM em MB
- *Tela*: Tamanho da tela em polegadas
- *CD*: Há um CD-ROM presente?
- *Multi*: Está incluído um kit multimídia (alto-falantes, placa de som)?
- *Premium*: O fabricante era uma empresa "premium" (IBM, COMPAQ)?
- *Anúncios*: Número de 486 listas de preços para cada mês
- *Tendência*: Tendência temporal indicando mês a partir de janeiro de 1993 a novembro de 1995
- *RAM*: A memória RAM é responsável pelo armazenamento de informações necessárias para a execução de aplicativos em uso e para o funcionamento do próprio sistema operacional.

Sendo assim, uma breve descrição do conjunto de dados onde temos, *Computers* são descrito em uma uma seção transversal de 1993 a 1995

- *número de observações*: 6259
- *observação*: Bens
- *país*: Estados Unidos

Pode ser consultado em data(computadores) no Shoftware *R*

Fonte: Stengos, T. e E. Zacharias (2005) "Preços intertemporais e discriminação de preços: uma análise hedônica semiparamétrica do mercado de computadores pessoais", Journal of Applied Econometrics, a ser publicado.

2. Análise Descritiva dos dados

Esta seção terá como objetivo a ser destacados algumas correlações entre as variáveis pressentes no banco de dados, por meio de uma breve análise descritiva, que podem auxiliar no caminho de um possível modelo em Regressão Linear Múltipla

A fim de determinar um modelo de Regressão Linear Múltipla que explica da melhor forma os preços em dólares americanos de 486 (PCs), ou seja, os computadores nos Estados Unidos no ano 1993 à 1995 feito por uma seção transversal, sendo assim, foi feito análises descritivas para entender como as demais variávies se distribuiem no banco de dados.

Para ilustrar o banco de dados, será feita uma análise descritiva de todas variáveis quantitativa. Na Figura 1, temos o gráfico de frequência da variável entre as variáveis *Preço em dolares Americano*, *Velocidade do clock em MHz*, *Tamanho do disco rígido em MB*, *A memória RAM*, *Tamanho da tela em polegadas*, *Número de 486 anúncios*, *Tendência temporal*.

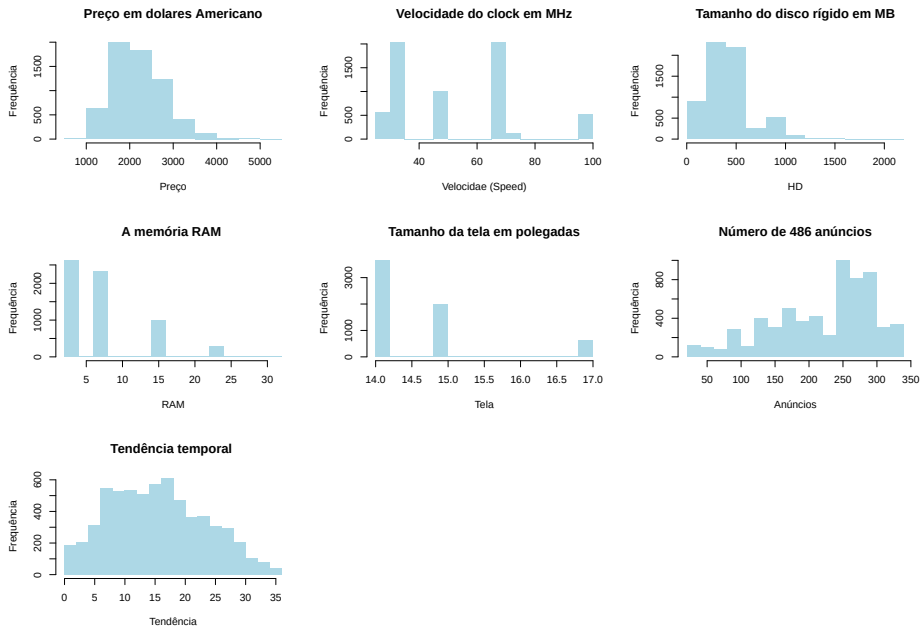

Figura 1: Gráficos das frequências de todas as variáveis quantitativa

Assim, ao fazer uma análise nos gráficos de frequências percebemos que na variável preço em dolares americano sua frequência chega em *1000* dolares, o mesmo a contede para cada uma das variávaies obtenha sua frequência máxima correspondente.

Sendo assim, ao analisar a Tabela 1 da Estatística descritiva das variáveis quantitativa dos cálculos de cada observação, bem como *mediana, média, desvio padrão, variância, amplitude, valores máximos e mínimos*. Percebemos que a variável *Price*, sendo o *Preço em Dolares Americnado* que está acima de todas as outras variáveis quando é feito a comparação com as estatísticas.

Table 1: Estatística Descritiva

	Price	Speed	HD	RAM	Screem	ABS	Trend
Média	2219.58	52.01	416.60	8.29	14.61	221.30	15.93
Mediana	2144.00	50.00	340.00	8.00	14.00	246.00	16.00
Disvio Padrão	580.80	21.16	258.55	5.63	0.91	74.84	7.87
Variância	337333.23	447.65	66847.30	31.71	0.82	5600.32	62.00
Amplitude	4450.00	75.00	2020.00	30.00	3.00	300.00	34.00
Máximo	5399.00	100.00	2100.00	32.00	17.00	339.00	35.00
Mínimo	949.00	25.00	80.00	2.00	14.00	39.00	1.00

É de se esperar que a variável Preço em Dolares Americano acaba sendo um valor alto em relação as outras variáveis, podemos perceber que o disvio padrão é *580.80*, o que segnifica que os preços dos computadores foram considerados em bem mais elevado do que o esperado.

Quando é feito as Estatísticas Descritiva de todas as variávies quantitativa na Tabela 1 do conjunto de dados, sendo assim, a variável *Price* tem um papel fundamental num possível modelo, chegando a ser a variável respota, com isso, será feito a correlação e comparação entre todas as outras observações nos gráficos. Afim de verificar o quanto de gasto estão ralacionado com as outras variáveis.

Ao examinar os dados da variável *Preço* vs *CD-ROM* na Figura 2, quanto foi feito a pergunta para obtermos um *Sim* ou *Não*, e para a variável qualitativa *CD-ROM*, nissso podemos perceber que na Figura 2, onde temos duas respostas da pesquisa de uma amostra aleátoria se no computador tem sua existencia do *CD-ROM*.

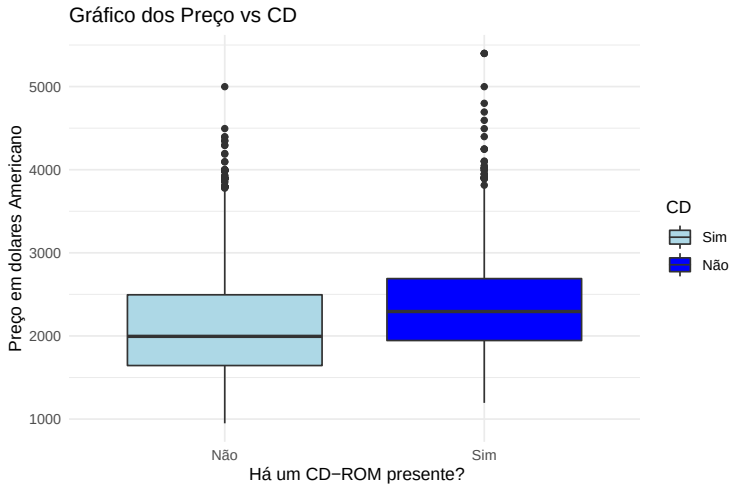

Figura 2: Gráfico de CD−ROM vs Preço em dolares

Quando comparando os Proços em dolares americas na pergunta se: *Há um CD-ROM presente?*, neste caso houver mais resposta para *Sim*, ou seja, nas lojas, fabrica tendoa sua saida de Computadores em uma grante desta amostra foram julgado como respsta *Sim*, desta maneira ao análisar o gráfico das respota *Sim* e *Não* tem mais ponto de outlirs acima da linha,ou seja, os outliers são dados que se diferenciam drasticamente de todos os outros, são pontos fora da curva normal *(*o que é curva normal?*)*. Em outras palavras, um outliers é um valor que foge da normalidade e que pode *(*e provavelmente irá*)* causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análies.

Por conseguinte, será analisado as duas últimas variáveis qualitativa sendo elas *Multi*, *Premium* na Figura 3 em relação ao preço em dolares americano, essas duas variáveis foram feito as perguntas com respostas *Sim* ou *Não*, para seguintes perguntas *Multi*: Está incluído um kit multimídia (alto-falantes, placa de som)?, *Premium*: O fabricante era uma empresa "premium" (IBM, COMPAQ)?. Ao fazer as comparações entre as duas perguntas nos dois gráficos de BoxPlots na Figura 3, percebe-se que para cada gráfico tem valores diferentes em suas respostas.

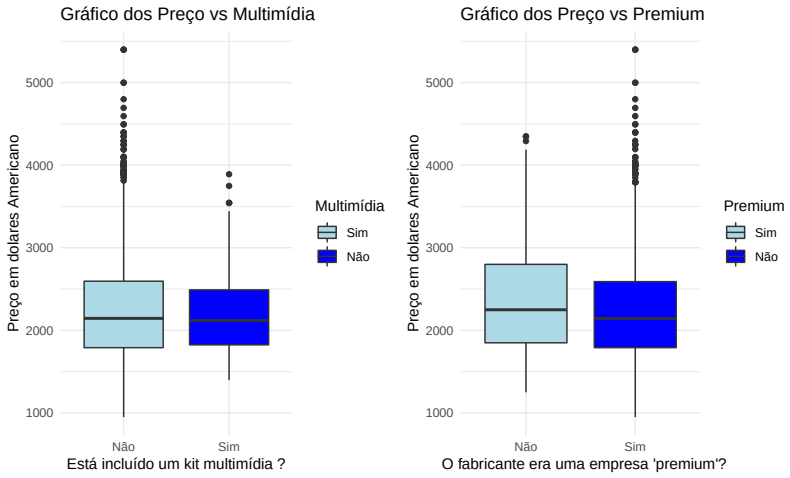

Figura 3: Gráficos de BoxPlots das perguntas

Desta maneira, quando a empresa de Computador do fabricante é considerado como *Premium* em suas respostas como *Sim* o preço em dolar americando fica acima 50000 considerando esses pontos como outilar, vale resaltar que acaba tendo uma grande quantidade de reposta para *Sim* para essa variável *Premium*, já

em relação a pergunta em *Multimídia* tem uma grande quantidade de respota que foram considerada como *Não*.

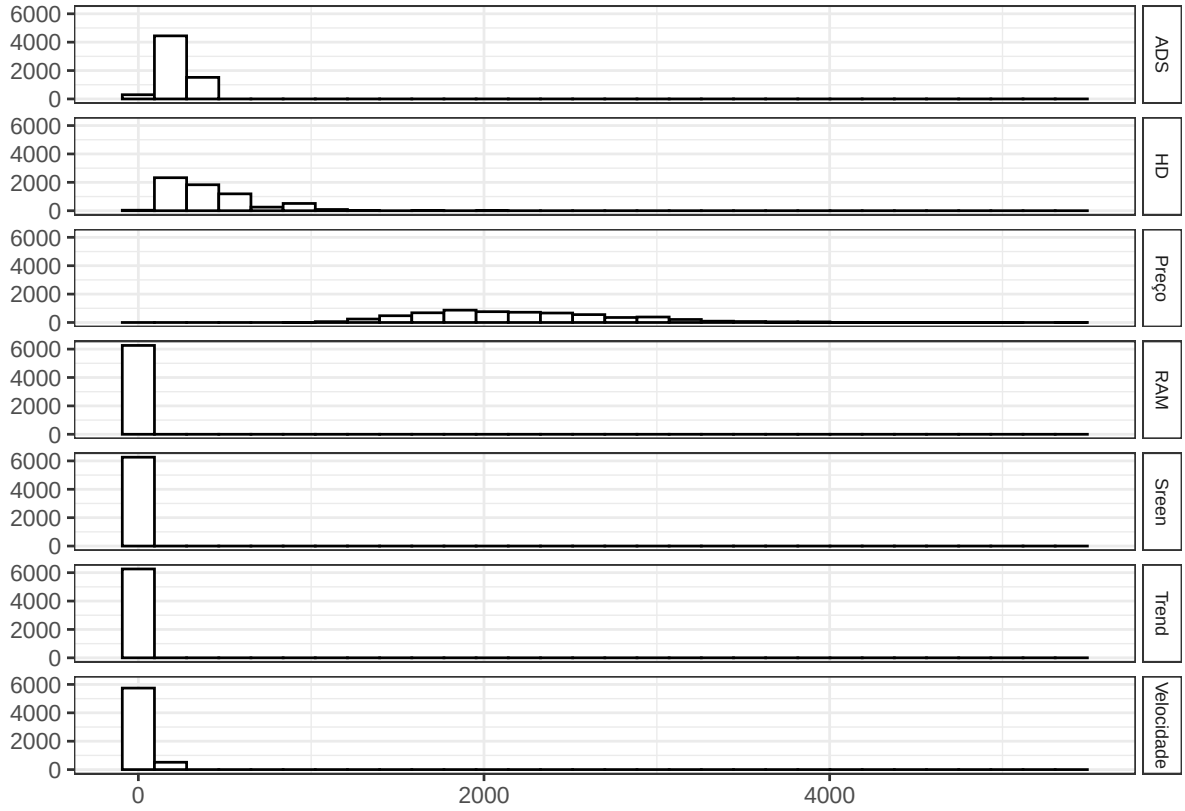

Figura 4: Gráficos de Barra

Logo, no gráfico de Barra da Figura 4 a variável *Preço* ela se distribui ao longo conjunto de dados, não se concentra soemtente um determinado intervalo, assim também na Tabela 1 faz um dialogo com a Figura 4, pois o *Preço em Dolares Americano* é uma variável importante nesta análise descritiva. Como as fabricas de computaroes tenta fazer os melhores produto de Computadores, o *Preço* acaba sendo uma variável resposta.

3. Suposição inferencial

Em diversos modelos teóricos existem uma relação de causalidade entre a variável-reposta com as demois variável regressora, esses fonôminos teóricos podem ser visto bem como em (antrópicos, biológicos, físicos, químicos, econômicos). Desta maneira, em muitos problemas práticos, o Modelo Regressão Linear Múltipla tem uma relação linear entre a resposta e a regressora é comprovada, mas o Modelo de Regressão Linear Simples não produz bons resultados pela omissão de regressoras importantes. A inclusão de nova(s) regressora(s):

- Eliminao o viés da equação de regressão;
- Reduz σ^2;
- Aumenta o poder dos testes;
- Reduz o comprimento dos IC's e IP's;
- Pode melhorar, piorar, simplificar ou complexar a interpretação do resultados;
- Pode tornar a coleta de dados mais cara e complexa

A técnica de Regressão Linear Múltipla pode ser usada para avaliar a relação entre um preditor e a reposta, enquanto "controlada" pelas demais variáveis no modelo

Modelo de regressão Linerar Múltipla geral

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_{2i} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

Onde:

- $\beta_0, \beta_1, ..., \beta_{p-1}$ são parâmetros.
- $X_{i1}, ..., X_{i,p-1}$ são constantes conheciddas.
- $\varepsilon_i \sim$ seguie uma i.i.d em $N(0, \sigma^2)$
- $i = 1, 2, ..., n$

Se $X_{i0} = 1$, podemos escrever:

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i$$

Devemos modelar a variável *Price* onde representa o *Preço em Dolares Americano* como resposta e as demais variáveis usadas como preditoras, propõe-se o desenvolvimento através de uma representação linear no seguinte formato em análise Modelo de Regressão Linear Múltipla:

$$Y_i = 307.98 + 9.32X_{i1} + 0.78X_{i2} + 48.25X_{i3} + 123.08X_{i4} + 60.91X_{i5} + 104.32X_{i6} - 509.22X_{i7} + 0.65X_{i8} - 51.84X_{i,9}$$

Neste modelo, existe todas as variáveis preditora presente no Modelo de Regressão Linear Múltipla, assim, o modelo proposto tem a seguinte variáveis onde os Y_i é o resposta e $\beta_i, i = 0, ..., 8$ são preditoras do modelo:

- Y_i é o *Preço*: Preço em dólares americanos de *486 PCs*
- β_0 é o *Velocidade*: Velocidade do clock em *MHz*
- β_1 é o *HD*: Tamanho do disco rígido em MB tamanho da memória *RAM* em *MB*
- β_2 é o *Tela*: Tamanho da tela em polegadas
- β_3 é o *CD*: Há um *CD-ROM* presente? *(1 para Sim,0 para Não)*
- β_4 é o *Multi*: Está incluído um kit multimídia *(alto-falantes, placa de som)*? *(1 para Sim,0 para Não)*
- β_5 é o *Premium*: O fabricante era uma empresa "premium" *(IBM, COMPAQ)*? *(1 para Sim,0 para Não)*
- β_6 é o *Anúncios*: Número de *486* listas de preços para cada mês
- β_7 é o *Tendência*: Tendência temporal indicando mês a partir de janeiro de *1993* a novembro de *1995*
- β_8 é o *RAM*: A memória *RAM* é responsável pelo armazenamento de informações necessárias para a execução de aplicativos em uso e para o funcionamento do próprio sistema operacional

Afim de análisar o modelo com todas as variáveis preditoras, devemos análisar o diagnósticos deste modelo, se existe homocedasticidade, variancia constante.

4. Diagnósticos do modelo

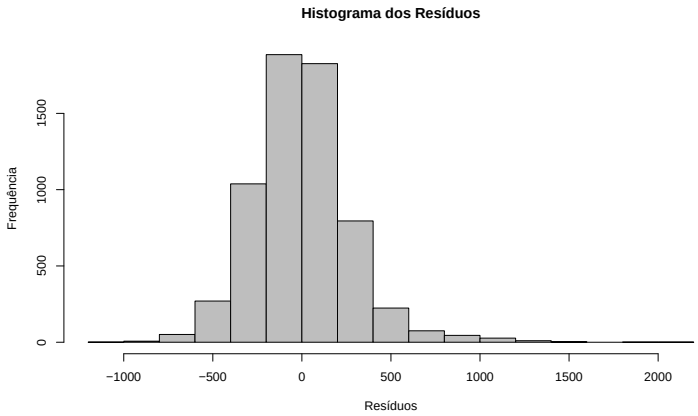

Figura 5: Gráficos de Histógrama dos Resíduos

Percebemos aqui,que existe uma centralidade nos Resíduos (Em termo Estatística os Resíduos é a diferença entre o valor observado e o valor real (ou mais próximo da realidade), numa variável; erro), quando análisando as caudas na Figura 5, não existe uma normalidade dos resíduos por quanta que as caudas ficam bem mais distribuidas para direita. Por isso será feito uma análise mais profunda na Figura 6, onde tem os 4 gráficos de regressão com os resíduos do modelo.

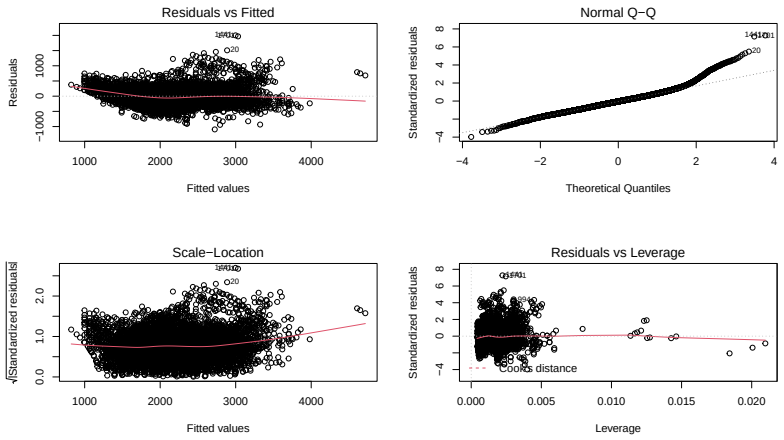

Figura 5: Gráficos de Histógrama dos Resíduos

O primeiro gráfico da Figura 6, *(Residula vs Fitted)* segue uma relação não linear entre os resíduos. O segundo, *QQplot*, explica uma possível normalidade dos erros, já que os pontos se distribuem quase uniformemente, mas quando análisando nas caudas não segue uniformemente no intervalo 4,-4 de acordo com a reta. O terceiro gráfico, de Locação Escala *(Scale-Location)* aponta para a heterocedasticidade dos resíduos, mas mesmo assim tem alguns pontos de aberrantes dos resíduos, e por fim, a última figura propõe visualmente uma variância constante entre os resíduos padronizados e aponta alguns pontos que podem alterar os resultados no modelo. Além disso, o modelo explica 77.56% da variância total dos dados, ou seja, $R^2 = 0.7756$.

5. Construção de um possível modelo

Nesta seção será construido um modelo que seja possível de interprestar e análisar. Para selecionar as variáveis que precisamos para um determinado modelo, usaremos o conceito de Cp de Mallows para ajudar a escolher entre modelos de regressao. Ele ajuda a obter um equilíbrio importante com o número de preditores no modelo. O Cp de mollows compara a precisão e o vício do modelo completo a models com um subconjunto de preditores.

Normalmente, poderiamos procurar modelos onde o Cp de Mallows é pequeno e está próximo do número de preditoras no modelo mais a constante *(p)*. Um pequeno valor de Cp de Mallows indica que o modelo é relativamente preciso *(tem pequena variância)* em estimar os coeficientes verdadeiros de regressão e predizer respostas futuras. Um valor de Cp de Mallows que está próximo do número de preditoras mais a constante indica que o modelo é relativamente não-viciado em estimar os coeficientes verdadeiros de regressão e predizer respostas futuras. Os modelos com falta de ajuste e vício têm valores de Cp de Mallows maiores do que *p*.

Proposta do modelo

Todas as variávies categóricas com p níveis diferentes foram categorizadas por meio de *p-1* níveis. A variável *CD*, por exemplo, foi categorizada por meio da variável "*Sim*", que recebeu "*1*" e se a resposta for "*Não*" é "*(zero) 0*" caso contrário. O mesmo aconteceu para as variáveis que recebeu as perguntas como *Sim* e *Não*.

Tendo em vista nisso, obtemos um banco de dados com as seguintes variáveis: *Preço*,*Velocidade*,*HD*, *Tela*,*CD*,*Multi*,*Premium*,*Anúncios*,*Tendência*,*RAM* Sendo assim a proposta do modelo ficara do seguinte formato:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_{2i} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, ..., 6259, \quad \beta_i, i = 1, .., 10$$

Explicação do módelo

A C_p de Mallow, para este critério avalia o erro quadrático méido dos n valores ajustados segundo um modelo a ser considerado.

Erro de cada valor ajustado é dados por:

$$\hat{Y}_i - \mu_i$$

em que μ_i é o valor verdadeiro da função resposta. Temos o viés:

$$E(\hat{Y}_i) - \mu_i$$

E um componente aleatório de erro:

$$\hat{Y}_i - E(\hat{Y}_i)$$

$$(\hat{Y}_i - \mu_i)^2 = [(E(\hat{Y}_i - \mu_i) + (\hat{Y}_i - E(\hat{Y}_i))]$$

$$E(\hat{Y}_i - \mu_i)^2 = [E(\hat{Y}_i - \mu_i] + Var(\hat{Y}_i)$$

Erro quadrático médio total:

$$\sum_{i=1}^{n} [E(\hat{Y}_i) - \mu_i]^2 + \sum_{i=1}^{n} Var(\hat{Y}_i)$$

Medida para o critério:

$$\Gamma_p = \frac{1}{\sigma^2} \left[\sum_{i=1}^{n} [E(\hat{Y}_i) - \mu_i]^2 + \sum_{i=1}^{n} Var(\hat{Y}_i) \right]$$

jugeito ao erro quadrático médio total dividido pela verdadeira variância do erro.

6. Medidas paliativas

A medida paliativa será em uma tomada de medida a ser tomada para um determinado modelo na construção do modelo de C_p de Mallow. Sendo assim, estamos considerando incluir *p-1* variáveis, mas assuma que o número ideal de variáveis a serem incluídas no modelos seja $P-1 > p-1$.

Se assumirmos que o modelo incluindo as $P-1$ variáveis é correto, temos que a Soma dos Quadrados Médios dos Erros, $QME(X_1, ..., X_{P-1})$ é um estimador não viesdo para σ^2.

Estimador para Γ_p é dado por:

$$C_p = \frac{SQE_p}{QME(X_1, ..., X_{P-1})} - (n - 2p)$$

7. Adequação do modelo inferencial

Como a base de dados é relativamente extensa, convém encontrarmos uma forma de filtrar e selecionar variáveis linearmente relevantes.

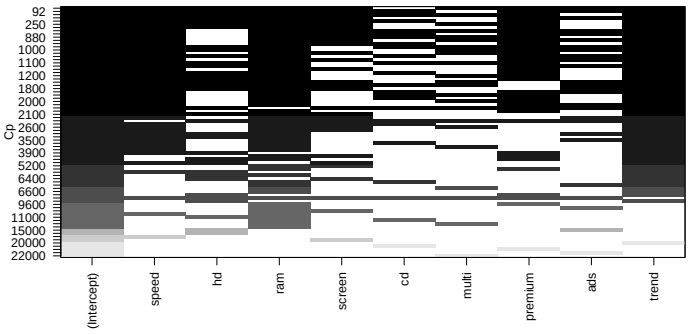

Figura 6: Modelos propostos pelo método C_p de Mallow I

Portando, temos as seguintes varívais:*Preço*, *Velocidade*, *HD*, *Tela*, *CD*, *Multi*, *Premium*, *Anúncios*, *Tendência*, *RAM*
Sendo assim a proposta do modelo I ficara do seguinte formato:

$Preço = \beta_0 + \beta_{1i} Velocidade + \beta_{2i} HD + Tela\beta_{3i} + CD\beta_{4i} + Multi\beta_{5i} + Premium\beta_{6i} + Anúncios\beta_{7i} + Tendência\beta_{8i} + RAM\beta_{9i}$

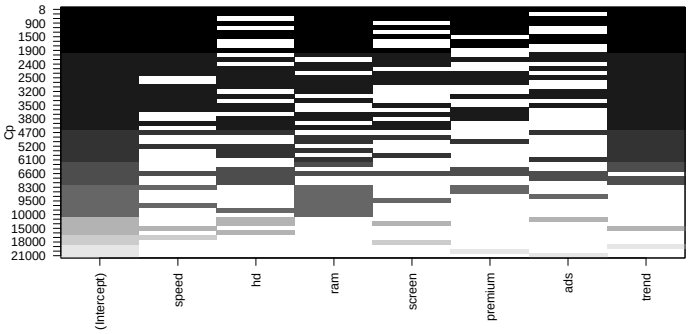

Figura 6: Modelos propostos pelo método C_p de Mallow II

De acordo com o modelo de C_p de Mallow, iremos retirar as variávies *CD* e *Multimídia*. Portando, modelo II ficara da seguinte forma:

$$Preço = \beta_0 + \beta_{1i}Velocidade + \beta_{2i}HD + Tela\beta_{3i} + Premium\beta_{6i} + Anúncios\beta_{7i} + Tendência\beta_{8i} + RAM\beta_{9i}$$

8. Analise dos Residuos e diagnósticos

Para verificar a normalidade, primeiro será observado o histograma dos resídos padronizados afim de si visualizar uma possível tendencia de uma normalidade quando retiramos as duas variávies *CD* e *Multimídia*.

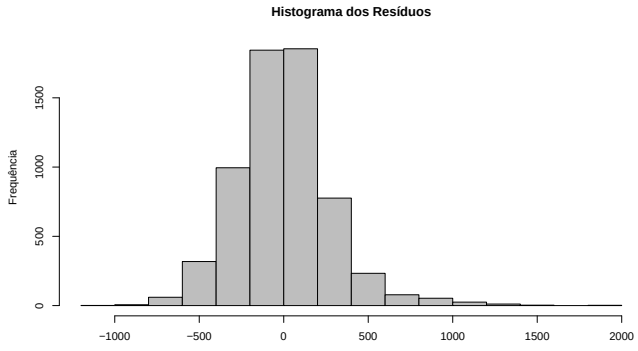

Figura 7: Distribuição do histogrma dos resíduos

Pelo o histograma acima nota-se descritivamente uma distribuição próxima da normal centrada no zero. Para comfirmar a normalidade será usado o gráfico Q-Q Plot e o teste de Lilliefors.

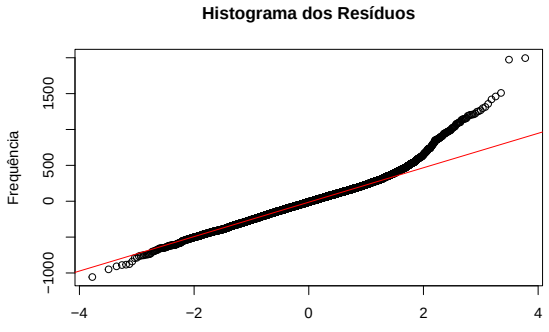

Figura 8: Distribuição do histograma dos resíduos

Graficamente, pode-se notar uma tendência de normalidade dos resíduos padronizados, porém, foi-se realizado todos os testes de hipótese nula não foi rejeitada nenhum com (p-valor $< \alpha$), ou seja, com base na amostra e com 95% de confiança, pode-se concluir que os resíduos padronizados seguem uma distribuição normal.

Para obtermos um gral de confiança do modelo qua vamos usar temos que testar a situação assumindo-se um Modelo de Regressão Linear Múltipla na situação em que $\beta_k = 0$. Nesta caso será feito o intervalo de confiança para os parâmetros de β_k, usaremos a distribuição t-student, assumindo

$$T = \frac{\beta_k}{\sqrt{Var(\hat{\beta}_k)}} \sim t_{48,2.5\%} = 2.010635$$

Table 2: Intervalos de Confiança para os parâmetros

	2.5 %	97.5 %
(β_0)	151.73	391.19
β_1	9.01	9.75
β_2	0.73	0.84
β_3	47.55	51.74
β_6	113.37	129.30
β_7	-501.94	-453.63
β_8	0.65	0.85
β_9	-50.54	-48.15

onde $IC(95\%, \beta_k) = (\beta_k \pm t_{48,2.5\%}\sqrt{\hat{Var}(\hat{\beta}_k)})$

Como nenhum dos intervalos contém o zero, então podemos atestar com 95% de confiança que β_k é não nulo, ou seja, as variáveis possuem de fato um significado no modelo. Vale ressaltar que os β_4 e β_5, não estão inclusa no modelo e também os X_{i4}, X_{i5}.

9. Interpretação do modelo e resultados e conclusões

Considerando o modelo II de Regressão Linear Múltipla, pelo métado de C_p de Mallow:

$$Preço = \beta_0 + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} + \beta_{3i}X_{3i} + \beta_{6i}X_{6i} + \beta_{7i}X_{7i} + \beta_{8i}X_{8i} + \beta_{9i}X_{9i}$$

O valor de β_0 é o intercepto do coeficiente ângular em 271.4559, ou seja, quando todas as outras variáveis forem igual a zero (0),sendo a esperança da resposta média estimada é em *Preço em Dolares Americando*, ou seja a variável (*Price*), neste modelo de Regressão Linear Múltipla pelo métado de C_p de *Mallow*, uma suposição é que se todos os outros β_{1i} onde $i = 1, .., 9$ forem iguais a zero (0), como não existe os β_{4i} e β_{5i}, então esse modelo é considerado como reduzido na falta de duas variável *CD* e *Múltimidia*.

O coeficiente β_1 indica o acréscimo de 9.377 na média final do *Preço em Dolares Americando* na resposta para a variável (*Preci*) quando é *Speed*, sendo a *Velocidade*, sendo assim, esse acréscimo írá acontecer quando forem mantidas as outras variáveis fixas no modelo.

O coeficiente β_2 indica um acréscimo de 0.7872 na média final do *Preço em Dolares Americando* na resposta para a variável (*Preci*) quando for *HD* e quando as outras variáveis estiverem fixas no modelo.

O coeficiente β_3 indica um acréscimo de 49.6473 na na média final do *Preço em Dolares Americando* na resposta para a variável (*Preci*) quando for RAM e quando as outras variáveis estiverem fixas no modelo.

O coeficiente β_6 indica um acréscimo de 121.3337 na média final do *Preço em Dolares Americando* na resposta para a variável (*Preci*) quando for *Screen*, ou seja, quando a variável é *Tela*, esse acréscimo írá acontecer quando forem mantidas as outras variáveis fixas no modelo.

O coeficiente β_7 indica um decréscimo de -477.7853 na média final do *Preço em Dolares Americando* na resposta para a variável (*Preci*) quando for *Premium* e quando as outras variáveis estiverem fixas no modelo.

O coeficiente β_8 indica um acréscimo de 0.7474 na média final do *Preço em Dolares Americando* na resposta para a variável (*Preci*) quando for *ads*, ou seja, quando a variável é *Anúncios*, esse acréscimo írá acontecer quando forem mantidas as outras variáveis fixas no modelo.

O coeficiente β_9 indica um decréscimo de -49.3413 na média final do *Preço em Dolares Americando* na resposta para a variável (*Preci*) quando for *trend*, ou seja, quando a variável é *Tendência*, esse acréscimo írá acontecer quando forem mantidas as outras variáveis fixas no modelo.

É importante ressaltar que o modelo prosposto para essa modelagem de dados explica apenas 76.81% da variação dos dados. No entando, modelos cujo propósito é predizer o comportantemento humando tendem a apresentar valores baixos para essa variável.

Referências Bibliográficas

https://cran.r-project.org/web/packages/Ecdat/Ecdat.pdf, *Acessada em 22 de Outubro de 2021*

https://weblogibc-co.com/wp-content/uploads/2018/07/Outliers-Malcolm-Gladwell.pdf, *Acessado em 26 de Outubro de 2021*

http://www.uesc.br/editora/livrosdigitais2/analiseexploratoria_r.pdf, *Acessado em 29 de Outubro de 2021*

http://cmq.esalq.usp.br/wiki/lib/exe/fetch.php?media=biometria:encontros:applied-regression.pdf, *Acessado em 31 de Outubro de 2021*