UNIVERSIDADE DE PASSO FUNDO
FACULDADE DE CIÊNCIAS ECONÔMICAS,
ADMINISTRATIVAS E CONTÁBEIS
CENTRO DE PESQUISA E EXTENSÃO DA FEAC
(www.upf.br/cepeac)

Texto para discussão

Texto para discussão Nº 07/2021

ANÁLISE DE REGRESSÃO LINEAR COM O AUXÍLIO DO RSTUDIO

Gustavo de Paula Pereira
Karina Deboni Grando
Matheus Machado

Passo Fundo - RS - Brasil

ANÁLISE DE REGRESSÃO LINEAR COM O AUXÍLIO DO RSTUDIO

Gustavo de Paula Pereira
Karina Deboni Grando
Matheus Machado

1 INTRODUÇÃO

A análise de dados voltada a compreensão de um fenômeno frente ao comportamento de uma ou mais variáveis potencialmente preditoras tem nas técnicas de regressão seu arcabouço teórico, trazendo este entendimento independentemente da relação de causa e efeito. Fávero e Belfiore (2017) complementa esta visão:

> "É de fundamental importância que o pesquisador seja bastante cuidadoso e criterioso ao interpretar os resultados de uma modelagem de regressão. A existência de um modelo de regressão não significa que ocorra, obrigatoriamente, relação de causa e efeito entre as variáveis consideradas".

Utilizadas massivamente em diversos setores do conhecimento, os modelos de regressão linear simples e múltipla possuem a capacidade, atendidos alguns pressupostos, de analisar os comportamentos de ativos no mercado financeiro, a variação de despesas em relação a expansão de uma indústria, ou até a influência da quantidade de quartos na formação do preço de venda em imóveis, entre outros (FÁVERO; BELFIORE, 2017).

Como forma de antecipar resultados em função de variáveis, simples ou múltiplas, a análise de regressão, ou regressão linear, busca gerar previsões testáveis das variáveis, diferentemente da correlação a qual não traz abordagens futuras sobre os elementos. Por meio das variáveis dependente (VD) e independentes (VIs), as VIs preveem valores de VD com auxílio do modelo preditivo, diferenciando a regressão simples da regressão múltipla pela quantidade de variáveis de saída, ou independente (FIELD, 2009).

2 REGRESSÃO LINEAR

Vistos como técnicas confirmatórias, os modelos de regressão devem ser utilizados com embasamento teórico subjacente, buscando "estimar o modelo desejado, analisar os resultados obtidos por meio de testes estatísticos e elaborar previsões" (FÁVERO; BELFIORE, 2017). Classificados como modelos lineares generalizados, os modelos de regressão linear trazem as características de (1) variável dependente quantitativa, (2) distribuição normal e (3) função de ligação canônica $\hat{Y}$.

Na visão de Field, Miles e Field (2012), por meio da equação 2.1, pode-se prever quaisquer dados:

$$resultado_i = (modelo) + erro_i \qquad 2.1$$

Considera-se que o resultado a ser buscado, por meio de qualquer modelo, possa ser previsto considerando a presença de um erro.

2.1 Regressão Linear Simples (RLS)

O modelo de regressão simples apresenta apenas uma variável *X*, representado pela função 2.1.1, onde podemos visualizar a disposição da equação no gráfico Figura 1, conforme Fávero e Belfiore (2017):

$$\hat{Y}_i = \alpha + \beta.X_i \qquad 2.1.1$$

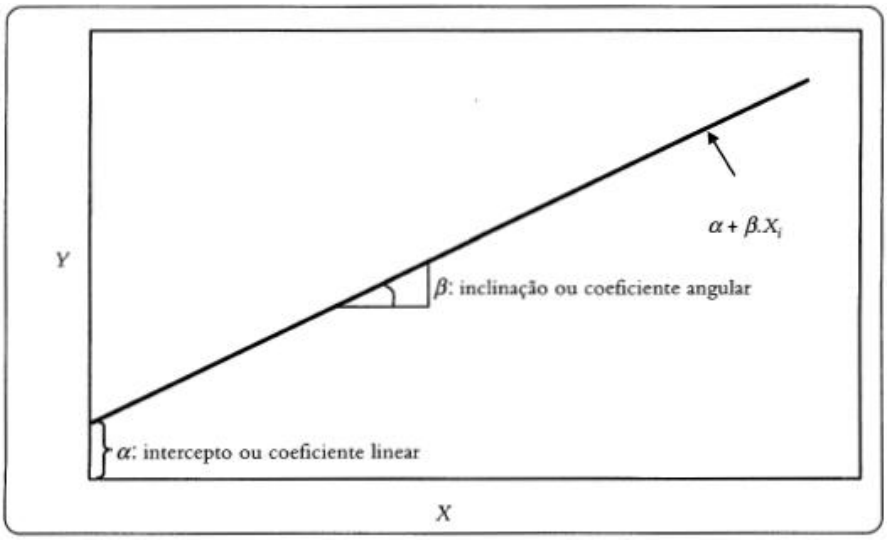

Figura 1 – Modelo estimado de regressão linear simples. Fávero e Belfiore (2017)

Podemos assim identificar o parâmetro α onde X = 0 frente a reta de regressão, chamado de intercepto, e o parâmetro β representando a inclinação da reta, sendo este atribuído pelo valor de relação com o acréscimo ou decréscimo de *Y* para cada *X*.

Em uma conversão da equação geral da tabela 2.1 ao ajuste da linha em função dos dados coletados, vemos os seguintes elementos: "(1) a inclinação (ou gradiente) da linha (geralmente denotada por b_1); e (2) o ponto em que a linha cruza o eixo vertical do gráfico (conhecido por *intercepto* da linha, b_0)" (FIELD; MILES; FIELD, 2012), além de (3) Y_i, resultado a ser previsto, e (4) X_i, a i^o pontuação sobre a variável de previsão, podendo ser representada da seguinte forma (2.2):

$$Y_i = b_0 + b_1X_i + \varepsilon_i \qquad 2.1.2$$

Sendo assim, sabendo o gradiente e o intercepto da linha, podemos descrever o modelo, e considerando no caso da regressão linear, tratar-se de uma linha reta, descrita matematicamente pela equação 2.2, podemos assim identificar a melhor linha atrelada aos dados, apurar o gradiente e o intercepto, e assim inserir variáveis preditoras no modelo para estimar variáveis resultados. "Basicamente então, o gradiente (b_1) nos diz a aparência do modelo (sua forma) e o intercepto (b_0) nos diz onde está o modelo (sua localização no espaço geométrico" (FIELD; MILES; FIELD, 2012).

O termo residual ε_i representa a margem de erro apresentado nos dados coletados, podendo na equação ser ignorado o termo, sendo a diferença entre a pontuação prevista para *i*, e a pontuação realmente obtida por *i*. Em outras palavras, significa que o "modelo não se ajustará perfeitamente aos dados coletados" (FIELD; MILES; FIELD, 2012). Para Fávero e Belfiore (2017), o termo de erro, ou residual, terá sua justificativa frente ao termo *Y*, onde

provavelmente apresentará relação com alguma variável *X* não incluída no modelo, e sendo assim, precisará ser representado pelo termo de erro, descrito por Fávero e Belfiore como *u*.

2.1.1 Mínimos Quadrados Ordinários

Por meio da regressão, os dados são dispostos em uma linha reta considerando ser linear. Esta linha pode ser traçada a olho nu, mas a subjetividade deste método exige uma forma mais precisa de apuração, sendo utilizado o método dos mínimos quadrados para melhor descrição dos dados (FIELD; MILES; FIELD, 2012).

Dentro de diversas possíveis linhas que podem ser traçadas buscando-se a menor quantidade de diferença entre os pontos, ainda haverá resíduos, e estes servem para avaliar o ajuste da linha de regressão. Medidos pela diferença vertical entre os dados e a linha, prevemos o valor de *Y* por meio da variável *X*, onde os pontos acima da reta mostram que o modelo subestima o valor, e quando abaixo, superestima-o (FIELD; MILES; FIELD, 2012).

Para Fávero e Belfiore (2017) devemos atender duas condições fundamentais relacionadas aos resíduos para estimarmos uma equação: (1) a somatória dos resíduos deve ser zero; porém somente esta condição ainda proporciona diversas retas de regressão, sendo assim, (2) a somatória dos resíduos ao quadrado é a mínima possível; desta forma, traz a reta que melhor se ajusta à nuvem de pontos, onde determina-se α *e* β de forma que "a somatória dos quadrados dos resíduos seja a menor possível" (FÁVERO; BELFIORE, 2017).

Para melhor apresentação teórica, utilizaremos o *software* RStudio e uma base de dados disponibilizada pelo pacote *ggplot2*, *mpg* – Dados de economia de combustível de 1999 a 2008 para 38 modelos populares de carros – considerando neste primeiro momento buscar por meio da regressão linear simples, saber a equação "milhas por galão (na estrada)" em função das "cilindradas do motor, em litros".

Podemos modelar o problema da seguinte forma:

$$milhas = f(cilindradas) \qquad 2.1.1.1$$

Sendo assim, o valor esperado da variável dependente para cada observação será:

$$\hat{milhas}_i = \alpha + \beta.cilindradas_i \qquad 2.1.1.2$$

Esta equação mostra o valor esperado para *milhas*, onde o mesmo é calculado para cada observação da amostra, em função da variável *cilindrada*, onde o *i* representa cada modelo de carro.

A base de dados utilizada foi nomeada como "*dados*" e a mesma se apresenta da seguinte forma (Figura 2), utilizando a fórmula *glimpse(dados)*, do pacote *dplyr*, onde "cilindradas do motor, em litros" equivale a coluna "*displ*" e "milhas por galão (na estrada)", equivalente a "*hwy*".

```
Rows: 234
Columns: 11
$ manufacturer <chr> "audi", "audi", "audi", "audi", …
$ model        <chr> "a4", "a4", "a4", "a4", "a4", "a…
$ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.…
$ year         <int> 1999, 1999, 2008, 2008, 1999, 19…
$ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4,…
$ trans        <chr> "auto(l5)", "manual(m5)", "manua…
$ drv          <chr> "f", "f", "f", "f", "f", "f", "f…
$ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, …
$ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, …
$ fl           <chr> "p", "p", "p", "p", "p", "p", "p…
$ class        <chr> "compact", "compact", "compact",…
```

Figura 2 – Apresentação visual da função *glimpse* no RStudio. O Autor

Por meio do gráfico de dispersão (Figura 3), vemos a relação Milhas por galão (na estrada) (Y) x Cilindradas do motor, em litros (X), onde cada ponto representa um modelo de veículo.

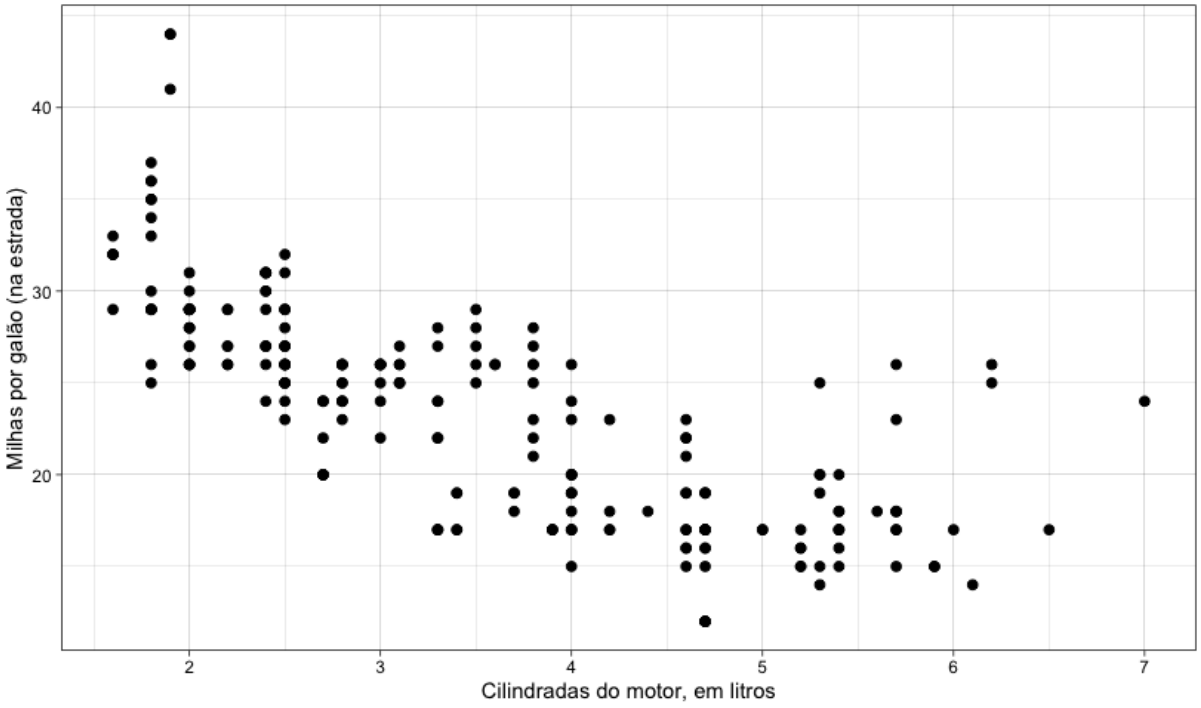

Figura 3 - Gráfico de dispersão Milhas por galão (na estrada) x Cilindradas do motor, em litros. O autor.

Para obter os coeficientes da reta de regressão linear, calculamos no RStudio por meio da função *summary(rls)* (Tabela 2.1.1.3) – O banco de dados "*rls*" foi criado para compilação de dados conforme imagem 3.1:

Call: 2.1.1.3

lm(formula = dados$hwy ~ dados$displ)

Residuals:

Min	1Q	Median	3Q	Max
-7.1039	-2.1646	-0.2242	2.0589	15.0105

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	35.6977	0.7204	49.55	<2e-16	***
dados$displ	-3.5306	0.1945	-18.15	<2e-16	***

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.836 on 232 degrees of freedom

Multiple R-squared: 0.5868, Adjusted R-squared: 0.585

F-statistic: 329.5 on 1 and 232 DF, p-value: < 2.2e-16

Sendo assim, podemos escrever a equação de regressão linear simples como, visto que por meio destes valores Coefficients > Estimate, podemos dizer que o intercepto ou Intercept (α) é 35,6977 e o coeficiente angular, representado por dados$displ, ($\beta$) é -3,5306:

$$Mîlhas_i = 35,6977 + (-3,5306).cilindradas_i \quad 2.1.1.4$$

Podemos alegar que os modelos que não possuem nenhuma cilindrada do motor, fazem 35,6977 milhas por litros (na estrada), não sendo este fato possível, considerando a análise física do fato, porém este fator pode ocorrer rotineiramente, sendo necessária uma apurada análise da parte do pesquisador e da teoria adjacente. Fisicamente seria impossível o deslocamento de um veículo sem cilindradas pela estrada, porém na visão matemática, este resultado reflete um prolongamento da reta até o eixo *Y* (FÁVERO; BELFIORE, 2012).

Interpretando a variável explicativa (*X*), a cada acréscimo nas cilindradas dos modelos de veículos da amostra reduz-se 3,5306 milhas por litros (na estrada) na autonomia dos veículos da relação. Exemplificando, um veículo com 1 litros a mais de cilindrada do motor, desempenha 3,5306 milhas a menos por litro na estrada do que outro sem este acréscimo. Por meio da Figura 4 vemos graficamente a reta de regressão linear simples do exemplo em questão:

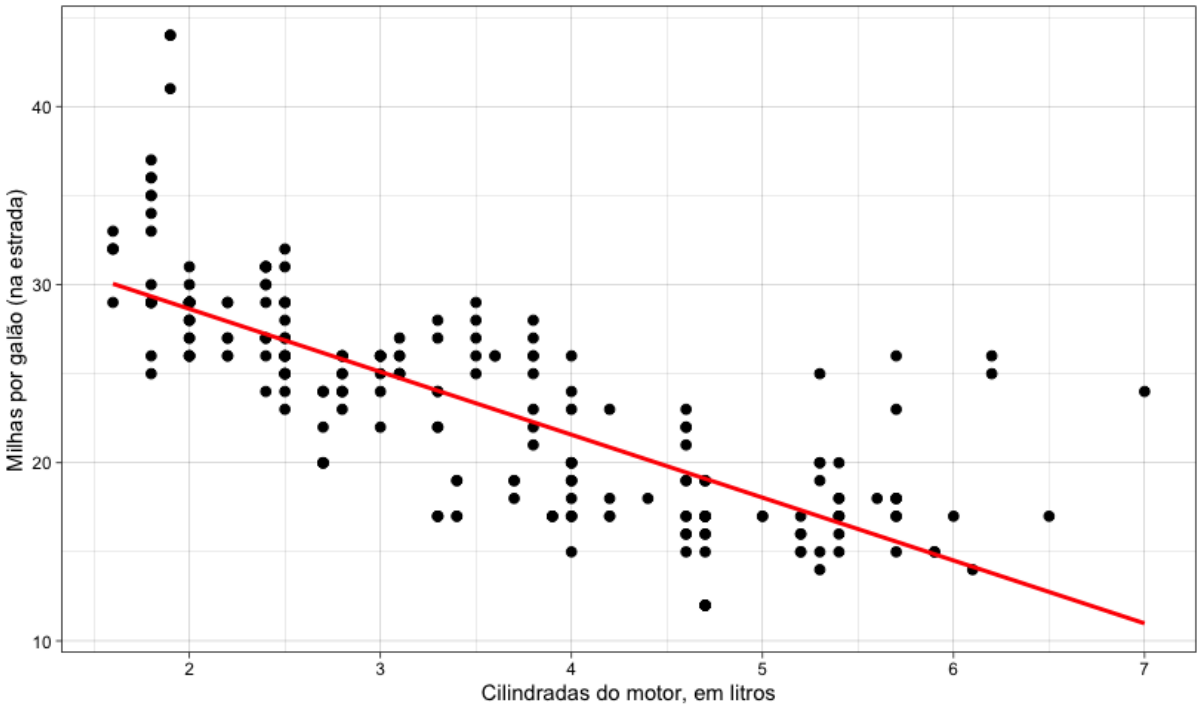

Figura 4 – Gráfico de dispersão com reta de regressão linear simples. O autor.

2.1.2 Coeficiente de Ajuste (R^2)

A qualidade do modelo deve ser avaliada buscando a otimização da apuração frente a média da amostra, o quão bem essa linha se ajusta aos dados reais, incorrendo uma comparação entre o modelo básico e o ajuste do melhor modelo (FIELD; MILES; FIELD, 2012):

$$desvio = \Sigma\ (observado - modelo)^2 \qquad 2.1.2.1$$

Para Fávero e Belfiore (2017), "o percentual de variabilidade da variável *Y* que é explicado pelo comportamento de variação das variáveis explicativas" podem ser mensuradas pelo coeficiente de ajuste R^2 onde:

> "a soma total dos quadrados (SQT) mostra a variação de *Y* em torno da própria média, a soma dos quadrados da regressão (SQR) oferece a variação de *Y* considerando as variáveis *X* utilizadas no modelo. Além disso, a soma dos quadrados dos resíduos (SQU) apresenta a variação de *Y* que não é explicada pelo modelo elaborado" (FÁVERO; BELFIORE, 2017).

Definido pela relação da tabela 2.1.2.2:

$$SQT = SQR + SQU \qquad 2.1.2.2$$

O modelo de regressão simples, por meio do R^2 da regressão apresenta "quanto do comportamento da variável *Y* é explicado pelo comportamento de variação da variável *X*, sempre lembrando que não existe, necessariamente, uma relação de causa e efeito entre as variáveis *X* e *Y*" (FÁVERO; BELFIORE, 2017). O R^2 obtém-se por meio da equação 2.1.2.3, podendo o resultado variar entre 0 e 1, sendo representado em forma de percentual (0% a 100%), sendo praticamente impossível a incidência de R^2 igual 1, este sendo 100%, visto que haveria a necessidade de todos os pontos se situarem em cima da reta, e a variabilidade de *Y*

estaria sendo totalmente explicada pelas variáveis de *X*. Em havendo R^2 igual a 0, a variação de *X* não corresponde a nenhuma variação em *Y* (FÁVERO; BELFIORE, 2017).

$$R^2 = SQR / SQT \qquad 2.1.2.3$$

Por meio tabela 2.1.1.3 identificamos o R^2 (Multiple R-squared), 0,5868, sendo possível afirmar que, 58,68% da variabilidade das milhas por galão (na estrada) é devido à variável da cilindrada do motor, em litros. E portanto, 41,32% desta variabilidade é devido a outras variáveis não incluídas no modelo, decorrentes da variação dos resíduos. Fávero e Belfiore (2017) ressalta que não se deve depositar demasiada confiança no coeficiente de ajuste R^2 visto que "a significância estatística geral do modelo e de seus parâmetros estimados não é dada pelo R^2, mas por meio de testes estatísticos apropriados" (FÁVERO; BELFIORE, 2017).

2.1.3 Significância Geral do Modelo

Como outra forma de utilização da soma dos quadrados, o teste *F* desempenha o papel de comparar o modelo com o erro no modelo, ou "a quantidade de variância sistemática dividida pela quantidade de variância assistemática" (FIELD; MILES; FIELD, 2012). Por meio desta medida podemos estimar o quanto melhorou o resultado frente ao nível de imprecisão do modelo, utilizando-se da fórmula 2.1.3.1, onde *k* representa o número de parâmetros do modelo (incluindo o intercepto) e *n*, o tamanho da amostra (FÁVERO; BELFIORE, 2017):

$$F = (SQR / (k - 1)) / (SQU / (n - k)) \qquad 2.1.3.1$$

Podemos observar, por meio da tabela 2.1.1.3, que o valor da estatística do teste (F-statistic) é 329,5 com 1 grau de liberdade na regressão e 232 para os resíduos, apresentando um nível de significância (p-value) de < 2.2e-16, este consideravelmente menor que 0,05 (ou 5%), parâmetro de nível de significância para identificação de hipótese nula ou alternativa.

Considerando as hipóteses nula e alternativa do teste *F*:

$$H_0: \beta = 0 \qquad 2.1.3.2$$
$$H_1: \beta \neq 0$$

Para darmos continuidade à análise de regressão, visto que o nível de significância ou p-value é menor que o parâmetro estipulado, valida-se a hipótese alternativa, sendo assim, $\beta_j \neq 0$. No modelo de regressão linear múltipla, em função de possuir mais de uma variável explicativa, onde o teste *F* avalia de forma conjunta a significância das variáveis, exige que o pesquisador avalie cada um dos parâmetros para inclusão no modelo (FÁVERO; BELFIORE, 2017).

Para Field, Miles e Field (2012), o teste *t* encontra-se por meio da identificação da razão da variância explicada pela variância inexplicada, ou erro. Buscando o valor de β diferente de zero por meio da equação 2.1.3.3, onde SE_β representa o erro padrão com o qual é associado, "o teste *t* nos diz se o valor β é diferente de 0 em relação à variação em *β-valores* nas amostras" (FIELD; MILES; FIELD, 2012).

$$t = \beta_{observado} / SE_\beta \qquad 2.1.3.3$$

O teste *t* por sua vez traz a verificação sobre a significância de cada parâmetro estimado, conforme tabela 2.1.3.4, podendo assim obter "valores críticos a um dado nível de significância e verificar se tais testes rejeitam ou não a hipótese nula" (FÁVERO; BELFIORE, 2017), utilizando-se de níveis de parâmetros semelhante ao teste *F*.

Se *valor-*P_t < 0,05 para o intercepto, $\alpha \neq 0$ 2.1.3.4

e

Se *valor -*P_t < 0,05 para determinada variável X, $\beta \neq 0$

Verificando-se diretamente o Pr(>|t|) na figura 2.1.1.3 podemos verificar se o nível de significância de cada variável é menor que 0,05 (5%) para validação do modelo. Para ambos os parâmetros temos o valor de <2e-16, e sendo assim, aceita como hipótese alternativa para as variáveis α e β. Fávero e Belfiore (2017) ressalta que:

> "o fato de não podermos rejeitar que o parâmetro α seja estatisticamente igual a zero a determinado nível de significância não implica que, necessariamente, devemos forçar a sua exclusão do modelo".

E complementa:

> "a não rejeição da hipótese nula para o parâmetro β a determinado nível de significância, por outro lado, deve indicar que a correspondente variável *X* não se correlaciona com a variável Y e, portanto, deve ser excluída do modelo final".

2.1.4 Intervalos de Confiança e Previsões

A construção de intervalos pelo RStudio se dá pelo comando *confint*, considerando a base de dados referência deste estudo, como visto na tabela 2.1.4.1, podendo assim selecionar o nível de confiança, ajustado para 95% devido padrão do estudo:

	2.5 %	97.5 %	2.1.4.1
(Intercept)	34.278353	37.11695	
dados$displ	-3.913828	-3.14735	

Onde vemos o intercepto, ou parâmetro α, representando seu intervalo de confiança por meio da equação (2.1.4.2):

$$P\ [34,278353 \leq \alpha \leq 37,11695] = 95\% \qquad 2.1.4.2$$

E o parâmetro β, sendo o coeficiente angular, visto na equação (2.1.4.3):

$$P\ [-3,913828 \leq \beta \leq -3,14735] = 95\% \qquad 2.1.4.3$$

Considerando que nenhum dos parâmetros englobam a hipótese de assumir o valor de 0, podemos rejeitar que sejam estatisticamente zero, conforme visto no teste *t*. Podendo-se interpretar a posição do parâmetro α no intervalo com 95% de confiança que se trata de valor

positivo, e este parâmetro encontra no intervalo [34,278353; 37,11695]. A mesma interpretação pode ser aplicada ao parâmetro β.

Por meio da aplicação do comando *confint(rls, level = 0.90)*, buscando a estimação do nível de confiança para 90%, identificamos no parâmetro α o intervalo [34,508001; 36,88730] e para o parâmetro β, [-3,851818; -3,20936], e sendo assim pode-se concluir que "quanto menores os níveis de confiança, mais estreitos (menor amplitude) serão os intervalos para conter determinado parâmetro" (FÁVERO; BELFIORE, 2017). Na figura 5 podemos ver a disposição gráfica do intervalo de confiança na reta de regressão.

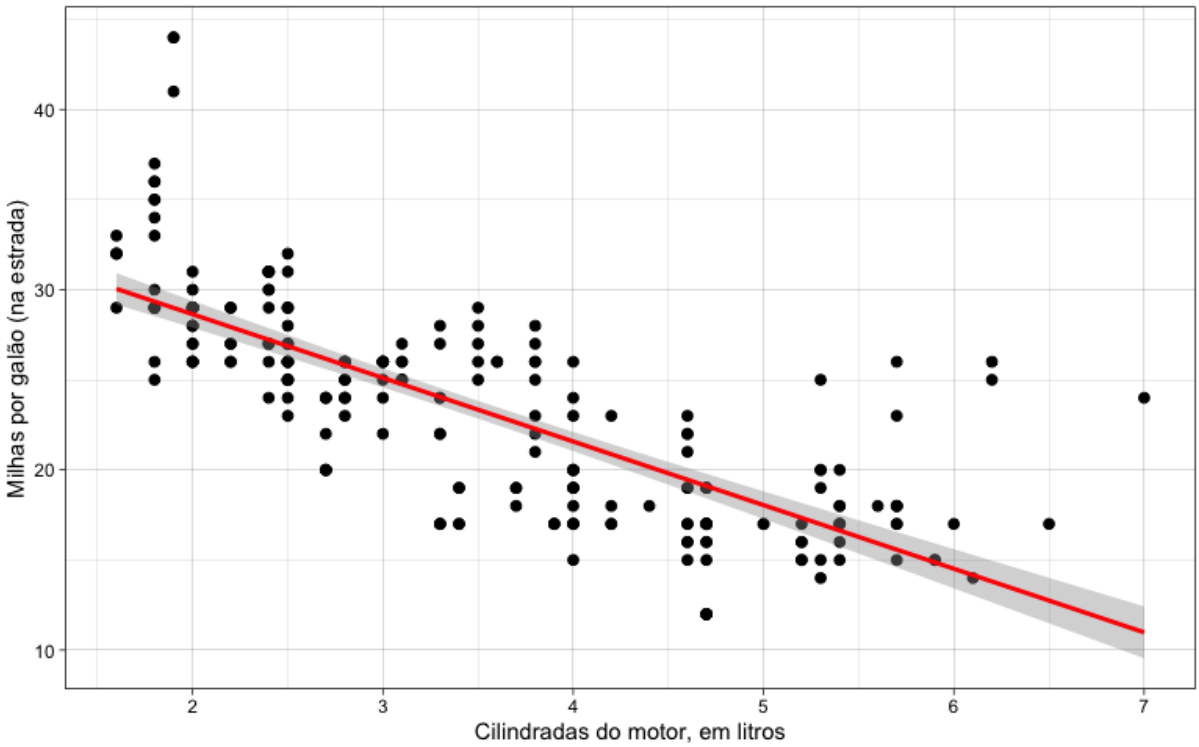

Figura 5 – Intervalo de confiança com reta de regressão linear simples. O autor.

Conforme Fávero e Belfiore (2017), a permanência do parâmetro β, onde variável X correlaciona-se com Y, deve-se os critérios de inclusão ou exclusão de parâmetros, conforme a figura 6. Tendo em vista a utilização do RStudio para este estudo, o critério Estatística *t* não foi abordado devido a necessidade de tabela complementar para análise do resultado, além da função exercida por esta análise ser completamente suprida pelo Teste *t*.

Parâmetro	Estatística *t* (para nível de significância α)	Teste *t* (análise do *valor-P* para nível de significância α)	Análise pelo Intervalo de Confiança	Decisão
β_j	$t_{cal} < t_{c\ \alpha/2}$	*valor-P* > nível de sig. α	O intervalo de confiança contém o zero	Excluir o parâmetro do modelo
	$t_{cal} > t_{c\ \alpha/2}$	*valor-P* < nível de sig. α	O intervalo de confiança não contém o zero	Manter o parâmetro no modelo

Obs.: O mais comum em ciências sociais aplicadas é a adoção do nível de significância α = 5%.

Figura 6 – Decisão de inclusão de parâmetros β_j em modelos de regressão (FÁVERO; BELFIORE, 2017).

Em uma abordagem exemplificada, buscando-se a quantidade de milhas por galão na estrada que um veículo com 3 litros de cilindradas faz, podemos efetuar as seguintes fórmulas considerando a apuração inicial, além dos parâmetros assumidos ao nível de confiança de 95% (2.1.4.4):

$$Mîlhas_i = 35,6977 + (-3,5306).cilindradas_i = 35,6977 + -3,5306.(3) = 25,1059\ milhas \quad 2.1.4.4$$

Tempo mínimo:

$$Mîlhas_{min} = 34,278353 + (-3,913828).cilindradas_i = 34,278353 + -3,913828.(3) = 22,5369\ milhas$$

Tempo máximo:

$$Mîlhas_{max} = 37,11695 + (-3,14735).cilindradas_i = 37,11695 + -3,14735.(3) = 27,6749\ milhas$$

Pode-se alegar então que, com nível de confiança de 95%, um veículo com 3 litros de cilindradas desempenha entre 22,5369 e 27,6749 milhas por galão, obtendo uma média 25,1059 milhas. Por fim, considerando que nesta amostra o R^2 foi de 0,5868, este indicador serve para mostrar a amplitude dos intervalos de confiança dos parâmetros, sendo assim, valores maiores que 0,5868 apresentariam nesta amostra uma precisão mais precisa, um gráfico de dispersão menos distante da reta de regressão, reduzindo assim a amplitude dos intervalos de confiança (FÁVERO; BELFIORE, 2017).

Considerando os pressupostos do modelo de regressão linear simples apresentados em comparação com os parâmetros de decisão de Fávero e Belfiore (2017), considera-se favorável a manutenção do modelo de $mîlhas_i = \alpha + \beta.cilindradas_i$ com nível de confiança de 95%.

2.2 Regressão Linear Múltipla (RLM)

Para Fávero e Belfiore (2017), o modelo geral de regressão linear traz a possibilidade de análise de estudo de uma ou mais variáveis explicativas, de forma linear, em função de uma variável dependente, apresentando-se da seguinte maneira:

$$Y_i = a + b_1.X_{1i} + b_2.X_{2i} + \ldots b_k.X_{ki} + u_i \quad 2.2.1$$

Classificando da seguinte forma: (1) Y traz a variável dependente quantitativa (fenômeno em estudo), (2) a representa o coeficiente linear ou intercepto, (3) $b_{1,2,3\ldots k}$ são os coeficientes angulares (coeficiente de cada variável), (4) $X_{1i,2i,3i\ldots ki}$ são as variáveis explicativas e (5) u é o termo de erro. Esta regressão traz a mesma lógica da regressão linear simples com o acréscimo de uma ou mais variáveis explicativas X, este dependendo da teoria subjacente e do estudo aplicado, levando sempre em consideração o bom senso do pesquisador.

A base de dados para este estudo será mantida a mesma da regressão linear simples, com a inclusão de uma variável, "quantidade de cilindros", constando na figura 2 como "*cyl*", apresentando assim uma equação de regressão múltipla no seguinte formato:

$$mîlhas_i = \alpha + \beta_1.cilindradas_i + \beta_2.cilindros_i \quad 2.2.2$$

Considerando que existem duas variáveis independentes no modelo em questão podemos representá-lo de forma gráfica por meio do gráfico de dispersão em três dimensões (Figura 7). Conforme Field (2009), na representação gráfica da RLM, o trapézio localizado no gráfico chama-se plano de regressão, onde "o plano ajustado aos dados pretende ser o melhor possível para esses dados. No entanto, existem invariavelmente algumas diferenças entre o modelo e os dados da vida real". Considerando a limitação do vinculada ao plano de três dimensões para a apresentação da regressão linear múltipla, ressalta-se que o RLM pode ser utilizada em três, quatros ou mais variáveis.

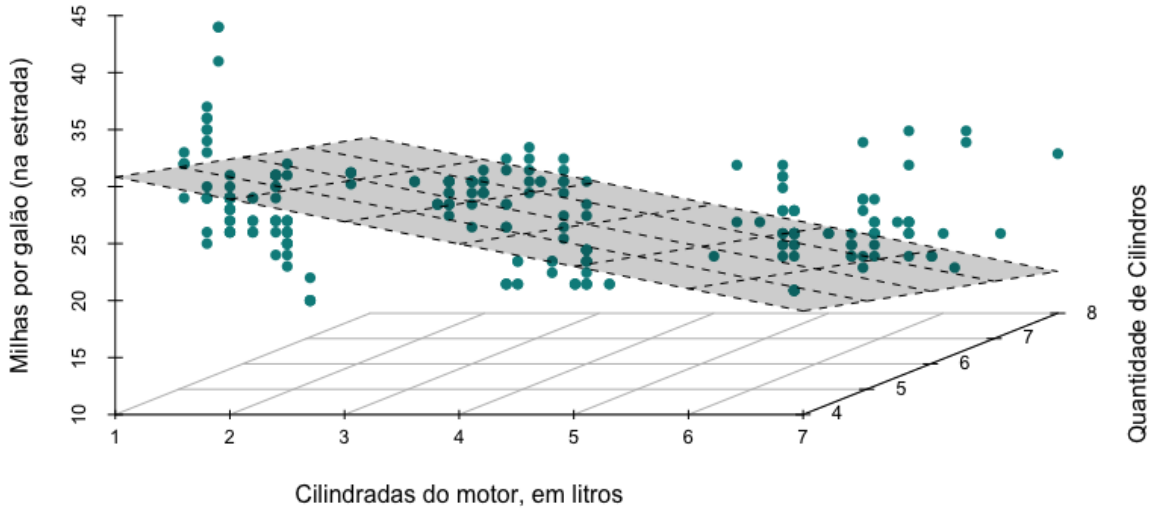

Figura 7 – Gráfico de dispersão das variáveis dispostas em 3D. O autor.

Semelhante ao comando utilizado para a identificação dos coeficientes da reta de regressão linear simples, a regressão linear múltipla apresenta-se no RStudio por meio da função *summary(rlm)* (Tabela 2.2.3) – O banco de dados "*rlm*" foi criado para compilação de dados conforme tabela 3.2:

Call: 2.2.3

lm(formula = dados$hwy ~ dados$displ + dados$cyl)

Residuals:

Min	1Q	Median	3Q	Max
-7.5098	-2.1953	-0.2049	1.9023	14.9223

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	38.2162	1.0481	36.461	<2e-16	***
dados$displ	-1.9599	0.5194	-3.773	0.000205	***
dados$cyl	-1.3537	0.4164	-3.251	0.001323	**

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.759 on 231 degrees of freedom

Multiple R-squared: 0.6049, Adjusted R-squared: 0.6014

F-statistic: 176.8 on 2 and 231 DF, p-value: < 2.2e-16

Sendo assim, a equação de regressão linear múltipla, conforme tabela 2.2.4, vista por meio dos valores Coefficients > Estimate, assume que o intercepto ou α (Intercept) é 38,2162, parâmetro β_1 (dados$displ) = -1,9599 e o parâmetro β_2 (dados$cyl) = -1,3537

$$Mîlhas_i = 35,6977 + (-1,9599).cilindradas_i + (-1,3537).cilindros_i \quad 2.2.4$$

Conforme o conceito de R^2, onde o apresenta o quanto da variância de *Y* justifica a regressão linear simples, o R^2 ajustado "nos informa quanta variância de Y pode ser creditada ao modelo se ele tiver sido derivado da população de onde a amostra for retirada" (FIELD, 2009). Para Fávero e Belfiore (2017), o conceito de R2 ajustado serve para comparar coeficientes de R2, visto sua capacidade de interpretar de modelos com tamanhos de amostra diferentes, ou quantidades diversas de parâmetros nos modelos, representada pela equação 2.2.5, onde o *n* representa o tamanho da amostra, e o *k* mostra a quantidade de parâmetros da amostra (variáveis explicativas mais intercepto):

$$R^2_{ajust} = 1 - ((n-1)/(n-k)*(1-R^2)) \quad 2.2.5$$

"O R^2 aumenta quando uma nova variável é adicionada ao modelo, entretanto o R^2 ajustado nem sempre aumentará, bem como poderá diminuir ou até ficar negativo", cita Fávero e Belfiore (2017). O valor do R2 ajustado, por meio do RStudio pode ser identificado na tabela 2.2.3 no campo *Adjusted R-squared: 0.6014*. Considerando o R2 ajustado da regressão linear simples, e comparação com a RLM, há um aumento nos valores, indicando a opção pelo segundo modelo ($R^2_{ajust,RLS}$ = 0,585 < $R^2_{ajust,RLM}$ = 0,6014).

A análise dos outputs deve ser executada também na regressão linear múltipla, onde se pode observar, por meio da tabela 2.2.3, que o valor da estatística do teste (F-statistic) é 176.8 com 2 graus de liberdade na regressão e 231 para os resíduos, apresentando um nível de significância (p-value) de < 2.2e-16, este consideravelmente menor que 0,05 (ou 5%), parâmetro de nível de significância para identificação de hipótese nula ou alternativa. Logo, pela análise de Pr(>|t|), identificamos que todos os parâmetros (α, β_1 e β_2) com os respectivos valores (<2e-16, 0,000205 e 0,001323) apresentam parâmetros válido para aceitação da hipótese alternativa para todas variáveis.

Os valores mínimos e máximos para o modelo em questão podem ser identificados no RStudio pelo comando *confint* com o nível de confiança, ajustado para 95% (2.2.6):

Tempo mínimo: 2.2.6

$$Mîlhas_i = 36,151002 + (-2,983299).cilindradas_i + (-2,174163).cilindros_i$$

Tempo máximo:

$$Mîlhas_i = 40,2813041 + (-0,93644445).cilindradas_i + (-0,5332101).cilindros_i$$

2.2.1 Variáveis *dummy*

A variável dummy, ou fictícia (FIELD, 2009) – Fávero e Belfiore (2017) utiliza o termo "binária" – visa representar grupos de pessoas por meio de zeros e um, criando tantas variáveis quantas forem necessárias considerando que a quantidade será uma a menos do que o número de categorias utilizadas.

A utilização arbitrária de valores para categorias qualitativas chama-se ponderação arbitrária, e caracteriza-se de um erro grave, onde se estaria "supondo que as diferenças na variável dependente seriam previamente conhecidas e de magnitudes iguais às diferenças dos valores atribuídos a cada uma das categorias da variável explicativa qualitativa" (FÁVERO; BELFIORE, 2017). Devendo estas variáveis serem utilizadas quando se busca a relação do comportamento de uma variável explicativa qualitativa e a variável dependente.

Para Field (2009), "o problema óbvio com a utilização de variáveis categóricas como previsores é que muitas vezes temos mais do que duas categorias". Por exemplo, uma mensuração de nível de satisfação dos consumidores de um tal produto, onde 0 refere-se a um consumidor totalmente insatisfeito e 5 onde o consumidor alega estar plenamente satisfeito, não podendo assim utilizar apenas a classificação de 0 ou 1.

Field (2009) utiliza uma classificação de oito passos para a análise das variáveis *dummy*, conforme tabela 2.2.1.1 adapta a classificação de Field visando esta análise:

1. Conte o número de grupos que você quer codificar e subtraia 1; 2.2.1.1
2. Crie tantas variáveis quanto o valor que foi determinado no passo 1. Essas são as variáveis *dummy*;
3. Escolha um dos seus grupos como base (isto é, o grupo contra o qual todos os demais serão comparados). Ele deve ser utilizado como grupo-controle, ou, se você não tem uma hipótese específica, ele deve ser o grupo que representa a maioria das pessoas (porque será importante comparar outros grupos contra a maioria);
4. Tendo escolhido o grupo-base, atribua a esse grupo o valor de 0 para todas as variáveis *dummy*;
5. Para a primeira variável *dummy*, atribua o valor de 1 ao primeiro grupo que você quer comparar contra o grupo-base. Atribua em todos os demais grupos o valor de 0 para essa variável;
6. Para a segunda variável *dummy* atribua o valor de 1 ao segundo grupo que você quer comparar ao grupo-base. Atribua a todos os demais grupos o valor 0 para essa variável;
7. Repita isso até que você não tenha mais variáveis *dummy*;
8. Coloque todas as variáveis *dummy* na análise de regressão.

Para a análise das variáveis *dummy*, utilizaremos o mesmo exemplo deste objeto de estudo com a inclusão da variável "*drv*", conforme figura 2, representando o tipo de tração dos veículos. Na amostra em questão, tendo em vista a variável "*dry*", por meio do comando *summary(dados$drv)*, podemos identificar a quantidade da amostra e os elementos f = tração nas rodas da frente, r = tração nas rodas traseiras, e 4 = tração nas quatro rodas.

4	f	r	2.2.1.2
103	106	25	

Considerando, na visão de Field (2009), em não havendo hipótese específica, deve-se selecionar o grupo com maior número de elementos, como o grupo-controle, sendo esta a variável "f", ou tração nas rodas da frente, e tendo em vista constarem 3 grupos, serão neste caso, 2 variáveis *dummy* – conforme RStudio, por meio da utilização da função *dummy_cols(dados$drv, remove_most_frequent_dummy = TRUE)*, *.data_4* e *.data_r*.

Sendo assim podemos escrever a seguinte equação de regressão múltipla considerando as variáveis *dummy* (2.2.1.3):

$$mîlhas_i = \alpha + \beta_1.cilindradas_i + \beta_2.cilindros_i + \beta_3.data\_4 + \beta_4.data\_r \quad 2.2.1.3$$

Por meio da tabela 2.2.1.4 podemos identificar os critérios para formação das variáveis *dummy* data_f e data_r, valores estes gerados pelo RStudio:

Categoria da variável qualitativa *dry*	Variável *dummy* data_f	Variável *dummy* data_r
f	0	0
4	1	0
r	0	1

2.2.1.4

Executando o comando *summary* novamente utilizando a base de dados *rlm.dummy* (criado conforme tabela 3.2 para compilação de dados) identificamos na tabela 2.2.1.5 os coeficientes da regressão linear múltipla com a inclusão de variáveis *dummy* em função da variável qualitativa "*dry*":

Call: 2.2.1.5

lm(formula = dados$hwy ~ dados$displ + dados$cyl + dados$drv$.data_4 + dados$drv$.data_r)

Residuals:

Min	1Q	Median	3Q	Max
-8.7095	-2.0282	-0.1297	1.3760	13.8110

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	38.1361	0.8407	45.360	<2e-16	***
dados$displ	-1.1245	0.4614	-2.437	0.0156	*
dados$cyl	-1.4526	0.3334	-4.357	1.99e-05	***
dadosdrv.data_4	-5.0446	0.5134	-9.826	< 2e-16	***
dadosdrv.data_r	-0.1595	0.8711	-0.183	0.8549	

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.968 on 229 degrees of freedom

Multiple R-squared: 0.7559, Adjusted R-squared: 0.7516

F-statistic: 177.2 on 4 and 229 DF, p-value: < 2.2e-16

Com isto, podemos identificar uma elevação considerada no R^2 ajustado, para 0,7516, porém o parâmetro *dadosdrv.data_r* na análise de nível de significância de 5% para diferença de zero não se mostrou diferente (Pr(>|t|) = 0,8549), sendo assim, esta variável será excluída e o modelo elaborado novamente, conforme tabela 2.2.1.6.

Call: 2.2.1.6

lm(formula = dados$hwy ~ dados$displ + dados$cyl + dados$drv$.data_4)

Residuals:

Min	1Q	Median	3Q	Max
-8.6760	-2.0152	-0.1351	1.4028	13.8132

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	38.1633	0.8257	46.217	<2e-16	***
dados$displ	-1.1612	0.4147	-2.800	0.00554	**
dados$cyl	-1.4425	0.3281	-4.396	1.68e-05	***
dadosdrv.data_4	-4.9904	0.4185	-11.925	< 2e-16	***

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.962 on 230 degrees of freedom

Multiple R-squared: 0.7558, Adjusted R-squared: 0.7526

F-statistic: 237.3 on 3 and 230 DF, p-value: < 2.2e-16

Conforme Fávero e Belfiore (2017), deve-se ter atenção a exclusão manual de variáveis visto que por meio da exclusão, os parâmetros os quais anteriormente constavam como estatisticamente diferente de zero, podem se tornar não diferente de zero, fato que não ocorre na amostra em questão. Sendo assim, podemos identificar o maior R^2 ajustado de todos os modelos analisados, sendo 0,7526, onde todos os parâmetros encontram-se diferente de zero, a um nível de significância de 5%, com o modelo final sendo:

$$Mîlhas_i = 38,1633 + (-1,1612).cilindradas_i + (-1,4425).cilindros_i + (-4,9904).data\_4 \quad 2.2.1.7$$

Exemplificando, quantas milhas por galão efetuaria um veículo com 3 cilindradas por litros possuindo 4 cilindros sendo um veículo de tração nas quatro rodas? A resposta seria 23,9193 milhas por galão, considerando a tabela 2.2.1.7, podendo interpretar que o aumento na quantidade de cilindros e cilindradas no veículo não afeta tanto a economia ou autonomia do veículo quanto o advento da tração traseira no mesmo. O valor atribuído de 1 ao parâmetro *data_4* advém da tabela 2.2.1.4 considerando que o modelo com maior significância foi o que excluiu o parâmetro *data_r*.

$$Mîlhas_i = 38,1633 + (-1,1612).3 + (-1,4425).4 + (-4,9904).1 \quad 2.2.1.7$$

2.2.2 Normalidade dos Resíduos

A normalidade dos resíduos serve para a validação dos valores de p-value nos testes *t* e teste *F*, assegurando assim os testes de hipótese dos modelos de regressão, porém ressalta-se que esta análise em grandes amostras pode minimizar uma violação. Recomenda-se a adequação deste pressuposto para "obtenção de uma série de resultados estatísticos voltados para a definição da melhor forma funcional do modelo e para a determinação dos intervalos de confiança para previsão (FÁVERO; BELFIORE, 2017).

Para Fávero e Belfiore (2017), visando a verificação dos pressupostos, recomenda-se a utilização dos testes Shapiro-Wilk, quando tratar-se de amostras com até 30 observações, e o teste Shapiro-Francia para grandes amostras. Considerando que a amostra utilizada neste estudo possui 234 observações, utilizaremos o teste de Shapiro-Francia por meio da função *sf.test(rlm.final$residuals)*, apresentando o resultado conforme tabela 2.2.2.1. Ressaltamos que o comando para a aplicação no RStudio para o teste de Shapiro-Wilk é *shapiro.test()*.

Shapiro-Francia normality test 2.2.2.1

data: rlm.final$residuals

W = 0.93393, p-value = 7.892e-08

Considerando o fato dos valores do modelo se mostrarem fora dos parâmetros adotado como distribuição normal (p > 0,05), classificado assim com representativo, este ressalta que a amostra difere de uma distribuição normal, porém, "um resultado significativo não necessariamente nos informa se o desvio da normalidade é suficiente para prejudicar os procedimentos estatísticos que serão aplicados aos dados" (FIELD, 2009).

2.2.3 Multicolinearidade

A multicolinearidade surge como problema quando há correlações elevadas entre as variáveis explicativas, indicando relação linear entre elas, surgindo principalmente quando apresentam tendências semelhantes, por exemplo, a utilização de índices de mercado atrelados a inflação como variáveis explicativas, onde muito provavelmente, ao longo do tempo, estes terão uma correlação, apresentando assim multicolinearidade. Outro fator possivelmente determinante para multicolinearidade é análise de amostra com poucas observações (FÁVERO; BELFIORE, 2017).

Para Fávero e Belfiore (2017), a multicolinearidade pode ser surgir de três formatos: correlação perfeita, valor de referência sendo 1, onde as observações possuem uma relação perfeita com mesmas características, não sendo possível calcular o parâmetro da variável quando ocorre; correlação muito alta, porém não perfeita, trazendo valores próximos de 1, e correlação baixa, onde a baixa correlação entre as variáveis gera pouca influência para a redução do teste *t*.

Para uma interpretação da multicolinearidade no RStudio, considerando a base de dados deste estudo, utiliza-se a função *pairs.panels(dados.rlm)* para a análise das correlações, conforme figura 8, onde vemos apenas na variável 2 (*dados$displ*) e 3 (*dados$cyl*) uma alta correlação, porém não sendo perfeita, estando menor que 1 (0,93 < 1).

Considerando Vasconcellos e Alves (2000) *apud* Fávero e Belfiore (2017):

> "A existência de multicolinearidade não afeta a intenção de elaboração de previsões, desde que as mesmas condições que geraram os resultados se mantenham na previsão. Desta forma, as previsões incorporarão o mesmo padrão de relação entre as variáveis explicativas, o que não representa problema algum".

Sendo assim, o importante com relação a multicolinearidade é "identificá-la, reconhecê-la e não fazer nada" (FÁVERO; BELFIORE, 2017).

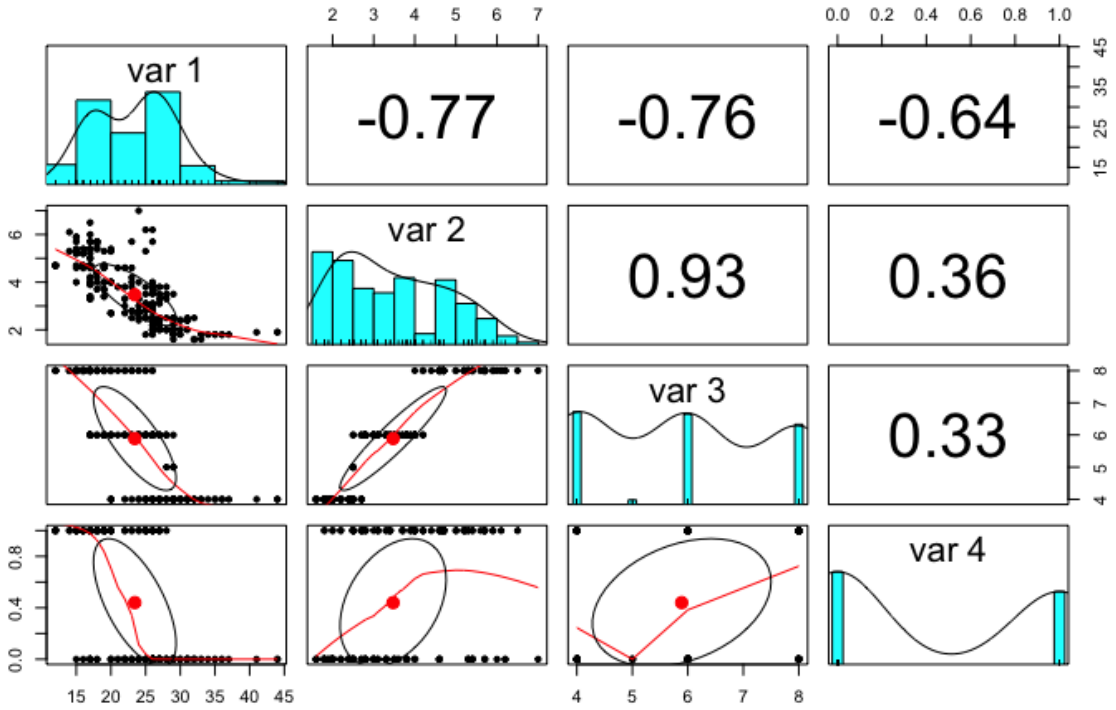

Figura 8 – Função *pairs.panel* no RStudio para análise de multicolinearidade. O autor.

Em complemento ao diagnóstico da multicolinearidade, derivado do resultado advindo da mesma, podemos identificar o VIF (*Variance Inflation Factor*), no RStudio por meio do comando *vif(rlm.final)*, onde os problemas de multicolinearidade podem ser identificado em parâmetros de VIF > 10, no caso deste estudo, identificamos, conforme tabela 2.2.3.1, que todas variáveis explicativas enquadradas nos parâmetros de VIF.

dados$displ	dados$cyl	var.dummy$.data_4	2.2.3.1
7.624035	7.428974	1.151367	

2.2.4 Heterocedasticidade

A análise da variância visa de forma uniforme suas distribuições perante cada termo, devendo assim, estas serem homocedásticas, onde quando da heterocedasticidade, "não há uma constância da variância dos resíduos ao longo da variável explicativa" (FÁVERO; BELFIORE, 2017). Podendo surgir devido erros no modelo, relações de aprendizagem e erro, além da presença de *outliers*, a heterocedasticidade acarreta problemas na hipótese de teste *t*.

Por meio do teste de Breusch-Pagan podemos identificar se a hipótese apresenta de forma constante a variância dos erros, sendo estes erros homocedásticos, ou como hipótese

alternativa, o fato da variância não ser constante, sendo esta verificação indicada para verificação da suposição de normalidade dos resíduos. Podemos assumir que se os termos forem homocedásticos, os resíduos aos quadrados não recebem a influência da variação de *Y*.

No software RStudio, o teste de Breusch-Pagan é efetuado pela fórmula *bptest(rlm.final)*, apresentando o seguinte resultado de 0,2215, mostrando um modelo homocedástico considerando que o mesmo supera o valor de 0,05.

studentized Breusch-Pagan test 2.2.4.1

data: rlm.final

BP = 4.3986, df = 3, p-value = 0.2215

2.2.5 Autocorrelação

A autocorrelação dos resíduos surge devido aos erros de especificação, ou omissão de variável explicativa relevante, além da influência sazonal, e estes podem ocasionar problemas no teste *t*. Para a identificação da presença de autocorrelação dos resíduos, aplicamos o teste Durbin-Watson. Após a identificação do problema, possíveis soluções serão a "alteração da forma funcional do modelo ou pela inclusão de variável relevante que havia sido omitida" (FÁVERO; BELFIORE, 2017).

Para a análise do problema da autocorrelação dos termos resíduos, o teste de Durbin-Watson verifica a incidência de correlação entre os erros, identificando se os resíduos adjacentes são correlacionados. Com parâmetros de 0 a 4, o teste de Durbin-Watson apresenta pela mediana 2 a não-existência de correlação, sendo que valores maiores que 2 trazem uma correlação negativa, e valores menores que 2, uma correlação positiva, conforme imagem 9. Conforme Field (2009), podem-se adotar os parâmetros entre 1 e 3, mantendo mesmo assim a abordagem conservadora do modelo.

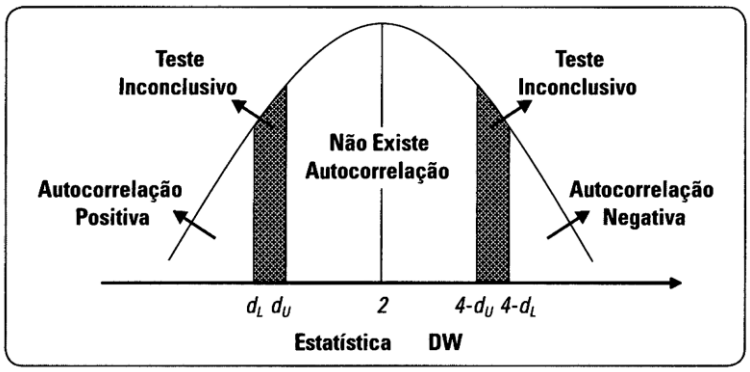

Figura 9 – Modelo estimado de regressão linear simples. Fávero e Belfiore (2017)

Por meio do comando *durbinWatsonTest(rlm.final)*, utilizando a base de dados vigente (2.2.5.1), vemos que a estatística Dubin-Watson (*D-W Statistic*) apresenta o valor de 1,273984 em sendo assim, de forma ainda conservadora, uma correlação positiva, porém não invalidando o modelo.

lag	Autocorrelation	D-W Statistic	p-value	2.2.5.1
1	0.3624753	1.273984	0	

Alternative hypothesis: rho != 0

2.3 Comparação entre Dois Modelos

Considerando a base de dados utilizada para o estudo, havendo-se identificado, após atendido os pressupostos para a regressão linear simples e a regressão linear múltipla, para Field, Miles e Field (2012), podemos comparar os modelos resultantes de ambos os estudos para a identificação do mais adequado. Por meio da função *anova(rls, rlm.final)*, somente de forma hierárquica, ou seja, o segundo modelo deve conter todos elementos do primeiro mais um elemento novo, e assim sucessivamente, pode identificar que há uma melhoria significativa do primeiro modelo em relação ao segundo considerando Pr(>F) < 0,05, no caso, < 2.2e-16.

Analysis of Variance Table 2.3.1

Model 1: dados$hwy ~ dados$displ

Model 2: dados$hwy ~ dados$displ + dados$cyl + dados$drv$.data_4

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	232	3413.8					
2	230	2017.3	2	1396.6	79.615	< 2.2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

3 EXECUTANDO NO RSTUDIO

O *software* utilizado RStudio será direcionado a resolução prática da teoria relacionada a análise da regressão linear simples e a regressão linear múltipla, não sendo abordadas questões introdutórias que visam apresentação prática da interface do *software*, instalação de pacotes, manipulação simples de dados e classes, restringindo-se aos comandos direcionados às técnicas confirmatórias.

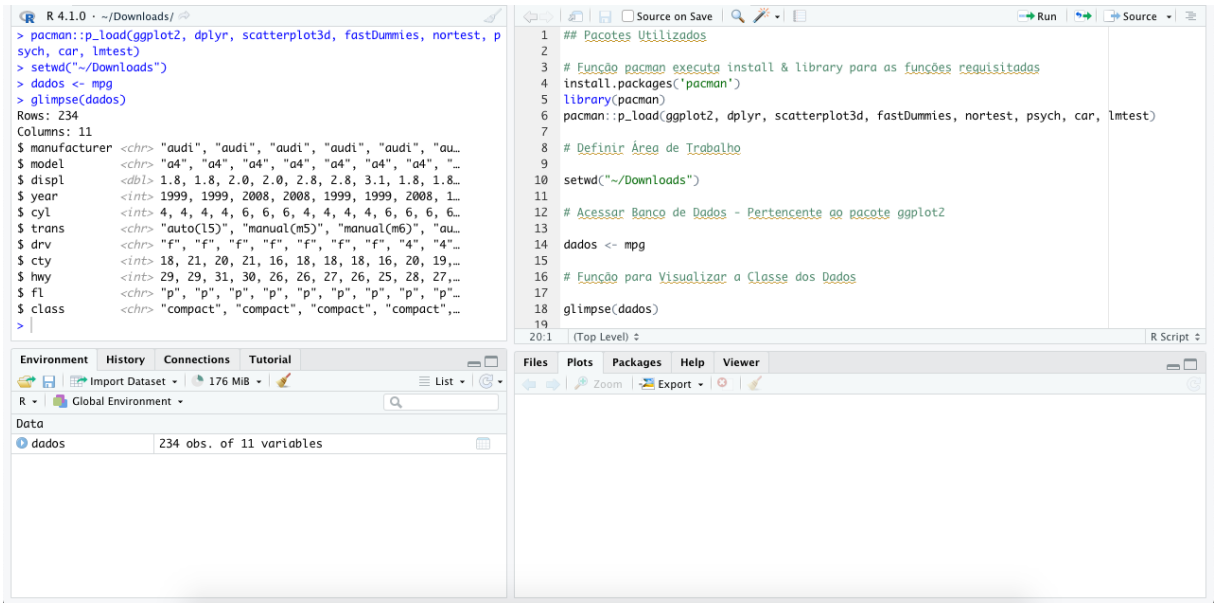

Figura 10 – Visualização do *software* RStudio. O autor.

O caractere # é utilizado no RStudio para inserir comentários não executáveis, sendo aplicado no estudo em questão para a explicação do comando. Conforme a figura 10, podemos identificar o resultado da utilização das funções primárias no RStudio visando a inserção e carregamento dos pacotes necessários, definição da área de trabalho, carregamento do banco de dados além da função *glimpse*, demonstrada na figura 2, para visualização do formato do conteúdo a ser utilizado.

A funções utilizadas para RLS neste trabalho serão listadas na íntegra na tabela 3.1 por ordem de apresentação:

ggplot(dados, aes(x=displ, y=hwy), size=3)+ geom_point(size=2)+ theme_linedraw() + labs(x = "Cilindradas do motor, em litros", y = "Milhas por galão (na estrada)")	Função: *ggplot* Pacote: *ggplot2* Finalidade: Representar graficamente dados em formato de dispersão Apresentação: Figura 3	3.1
rls<-lm(dados$hwy ~ dados$displ)	Função: *lm* Pacote: *base* Finalidade: Utilizada para apresentação de regressão linear Apresentação: Compilação de dados para utilização futura	
summary(rls)	Função: *summary* Pacote: *base* Finalidade: Apresentar um resumo dos dados de referência Apresentação: Tabela 2.1.1.3	
ggplot(dados, aes(x=displ, y=hwy), size=3)+ geom_point(size=2)+ geom_smooth(method=lm, col="red", se=FALSE)+ theme_linedraw() + labs(x = "Cilindradas do motor, em litros", y = "Milhas por galão (na estrada)")	Função: *ggplot* Pacote: *ggplot2* Finalidade: Representar graficamente dados em formato de dispersão com a inclusão da reta de regressão linear Apresentação: Figura 4	
confint(rls)	Função: *confint* Pacote: *base* Finalidade: Apresentar os coeficientes a um nível de confiança de 95% Apresentação: Tabela 2.1.4.1	
confint(rls, level=0.9)	Função: *confint*	

	Pacote: *base* Finalidade: Apresentar os coeficientes a um nível de confiança de 90% Apresentação: Dados de referência
ggplot(dados, aes(x=displ, y=hwy), size=3)+ geom_point(size=2)+ geom_smooth(method=lm, col="red", se=TRUE)+ theme_linedraw() + labs(x = "Cilindradas do motor, em litros", y = "Milhas por galão (na estrada)")	Função: *ggplot* Pacote: *ggplot2* Finalidade: Representar graficamente dados em formato de dispersão com a inclusão da reta de regressão linear e as margens do nível de confiança a 95% Apresentação: Figura 5

A funções utilizadas para RLM neste trabalho serão listadas na íntegra na tabela 3.2 por ordem de apresentação:

rlm <- lm(dados$hwy ~ dados$displ + dados$cyl)	Função: *lm* Pacote: *base* Finalidade: Utilizada para apresentação de regressão linear Apresentação: Compilação de dados para utilização futura	3.2
graph <- scatterplot3d(dados$hwy ~ dados$displ + dados$cyl, pch=16, color = "darkcyan" , box=FALSE, zlab = "Milhas por galão (na estrada)", xlab = "Cilindradas do motor, em litros", ylab = "Quantidade de Cilindros") graph$plane3d(rlm, col="black", draw_polygon = TRUE)	Função: *scatterplot3d* Pacote: *scatterplot3d* Finalidade: Representar graficamente em 3D dados em formato de dispersão com plano gráfico da regressão linear múltipla Apresentação: Figura 7	
summary(rlm)	Função: *summary* Pacote: *base* Finalidade: Apresentar um resumo dos dados de referência Apresentação: Tabela 2.2.3	
confint(rlm)	Função: *confint* Pacote: *base* Finalidade: Apresentar os coeficientes a um nível de confiança de 95% Apresentação: Tabela 2.2.6	

dados$drv<-as.factor(dados$drv)	Função: *as.factor* Pacote: *base* Finalidade: Conforme imagem 1, converter os dados de character para *factor*, para contagem por meio de *summary* Apresentação: Compilação de dados para utilização futura
summary(dados$drv)	Função: *summary* Pacote: *base* Finalidade: Apresentar um resumo dos dados de referência Apresentação: Tabela 2.2.1.2
dummy_cols(dados$drv, remove_most_frequent_dummy = TRUE)	Função: *dummy_cols* Pacote: *fastDummies* Finalidade: Criar uma tabela binária para variáveis qualitativas Apresentação: Compilação de dados para utilização futura
rlm.dummy <- lm(dados$hwy ~ dados$displ + dados$cyl + dados$drv$.data_4+ dados$drv$.data_r)	Função: *lm* Pacote: *base* Finalidade: Utilizada para apresentação de regressão linear Apresentação: Compilação de dados para utilização futura
summary(rlm.dummy)	Função: *summary* Pacote: *base* Finalidade: Apresentar um resumo dos dados de referência Apresentação: Tabela 2.2.1.5
rlm.final <- lm(dados$hwy ~ dados$displ+ dados$cyl + dados$drv$.data_4)	Função: *lm* Pacote: *base* Finalidade: Utilizada para apresentação de regressão linear Apresentação: Compilação de dados para utilização futura
summary(rlm.final)	Função: *summary* Pacote: *base* Finalidade: Apresentar um resumo dos dados de referência Apresentação: Tabela 2.2.1.6
sf.test(rlm.final$residuals)	Função: *sf.test* Pacote: *nortest* Finalidade: Apresentar o resultado da amostra aplicada ao teste Shapiro-Francia Apresentação: Tabela 2.2.2.1
dados.rlm<-cbind(dados$hwy, dados$displ, dados$cyl,	Função: *cbind* Pacote: *base*

dadosdrv.data_4)	Finalidade: Compilação dos dados da planilha base para análise de multicolinearidade Apresentação: Compilação de dados para utilização futura
pairs.panels(dados.rlm)	Função: *pairs.panel* Pacote: *psych* Finalidade: Apresentação dos dados RLM para análise de multicolinearidade Apresentação: Figura 8
vif(rlm.final)	Função: *vif* Pacote: *car* Finalidade: Diagnóstico para análise de multicolinearidade Apresentação: Tabela 2.2.3.1
bptest(rlm.final)	Função: *bptest* Pacote: *lmtest* Finalidade: Apresentar o resultado da amostra aplicada ao teste Bresuch-Pagan Apresentação: Tabela 2.2.4.1
durbinWatsonTest(rlm.final)	Função: *durbinWatsonTest* Pacote: *car* Finalidade: Apresentar o resultado da amostra aplicada ao teste Durbin-Watson Apresentação: Tabela 2.2.5.1
anova(rls, rlm.final)	Função: *anova* Pacote: *base* Finalidade: Verificar a análise de variância entre um ou mais modelos Apresentação: Tabela 2.3.1

4 REFERÊNCIAS

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados**. Rio de Janeiro: Elsevier, 2017.

FIELD, A. **Descobrindo a estatística usando o SPSS**. Porto Alegre, Artmed, 2009.

FIELD, A. P.; MILES, J.; FIELD, Z. **Discovering statistics using R**. SAGE Publications, 2012.