Predição de consumo de carros

Universidade de Brasília

Instituto de Ciências Exatas - Departamento de Estatística

Análise de Regressão Linear

Alunos Responsáveis:

Álvaro Jeronimo da Silva Kothe - 17/0004694

Marcos Augusto Daza Barbosa - 17/0017834

Mathews de Noronha Silveira Lisboa - 17/0042324

William Edward Rappel de Amorim - 17/0047385

Professora:

Juliana Betini Fachini Gomes

Brasília, 9 de Julho de 2019

Conteúdo

1 Introdução

O seguinte projeto tem como objetivo estudar e predizer o consumo de combustível em milhas por galão de automóveis no ano de 1983. Com este intuito, serão realizadas análises descritivas exploratórias para elencar os principais fatores que devem influenciar o consumo do carro, para em seguida ajustar o modelo de regressão linear mais adequado aos dados.

O banco de dados possui 398 observações, cada uma relativa a um automóvel diferente, contendo as seguintes informações:

- *Car name*: marca e modelo;
- *Mpg*: consumo em milhas por galão;
- *Cylinders*: número de cilindros do motor;
- *Displacement*: deslocamento do motor em polegada cúbica, que equivale a cilindrada;
- *Horsepower*: potência em cavalos;
- *Weight*: peso do veículo em libras;
- *Acceleration*: tempo em segundos para aceleração de 0 a 60 milhas por hora;
- *Age*: idade no carro no ano de 1983;
- *Origin*: região de origem (Estados Unidos, Europa ou Japão).

O software utilizado para elaboração de tabelas, gráficos e análises foi o R versão 3.4.3.

2 Metodologia

2.1 Regressão linear múltipla

O modelo de regressão linear múltiplo de ordem 1 com p variáveis tem a forma

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \tag{1}$$

- Y_i é o valor da variável resposta para a i-ésima observação;
- $\beta_0, \beta_1, \ldots, \beta_p$ são parâmetros desconhecidos;
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ são constantes conhecidas;
- ε_i são independentes e $\mathcal{N}(0, \sigma^2)$.

A função resposta para o modelo é

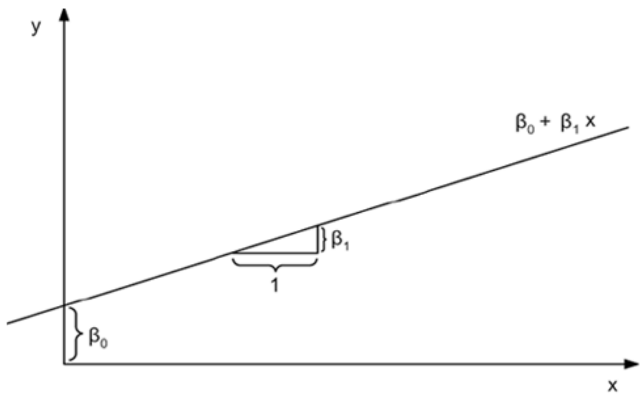
$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{2}$$

Quando $p = 1$ o modelo é um modelo linear simples, da forma

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \tag{3}$$

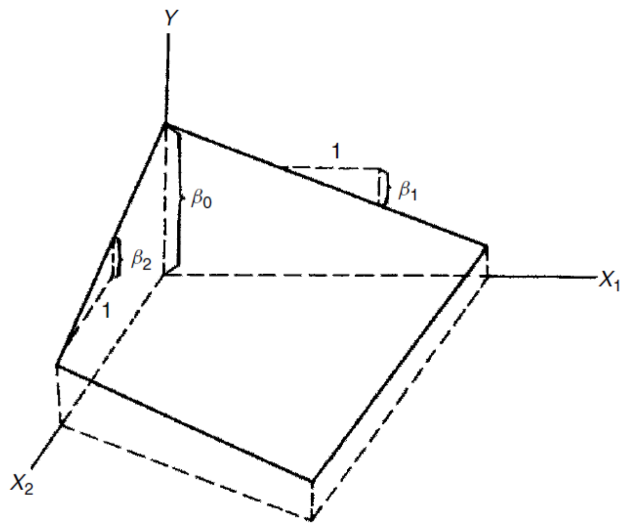
em que o seu gráfico é:

Figura 1: Modelo de regressão linear simples



Outro caso especial é quando $p = 2$, o que traz uma superfície de resposta na forma

Figura 2: Superfície de regressão com 2 variáveis



E para estimar $\beta_0, \beta_1, \ldots, \beta_p$ é usado mínimos quadrados ordinários, em que as estimativas de Beta são:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{4}$$

Em que:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \tag{5}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \tag{6}$$

2.2 Valor predito

O valor predito é o valor que o modelo prediz, considerando os valores das variáveis explicativas, que é definido como:

$$\mathbf{Y} = \mathbf{X}\hat{\beta} \tag{7}$$

2.2.1 Intervalo de Predição

Considerando $\hat{Y}_n$ o valor predito de uma nova variável, a variável $\hat{Y}_n$ segue uma distribuição Normal com média $E(Y_n)$ e variância $\sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$. Com isso, um intervalo de predição $(1-\alpha)$ para $E(Y_n)$ é dado por:

$$IC(E(Y_n); 1-\alpha) = \left(\hat{Y}_n - t_{(1-\frac{\alpha}{2}; n-(p+1))}\sqrt{QMResA}; \hat{Y}_n + t_{(1-\frac{\alpha}{2}; n-(p+1))}\sqrt{QMResA}\right) \quad (8)$$

Em que $A = \left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]$.

2.3 Resíduo

Resíduo é a estimativa do erro do modelo, dado pela diferença entre o valor observado e o predito, definido como

$$e_i = Y_i - \hat{Y}_i \quad (9)$$

2.4 Soma de Quadrados

É importante definir as somas de quadrados para resíduos, totais e regressão. Assim como seus respectivos quadrados médios. Sendo assim para soma de quadrados de resíduo define-se:

$$SQRes = \sum_{i=1}^{n} e_i^2 \quad (10)$$

$$QMRes = \frac{SQRes}{G.L(resduos)} \quad (11)$$

Em que G.L são os graus de liberdade do resíduos.

Para soma de quadrados de regressão define-se:

$$SQReg = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \quad (12)$$

$$QMReg = \frac{SQReg}{G.L(regresso)} \quad (13)$$

Sendo que G.L são os graus de liberdade da regressão.

Para soma de quadrados totais define-se:

$$SQTotal = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{14}$$

$$QMTotal = \frac{SQTotal}{G.L(Totais)} \tag{15}$$

Sendo que G.L são os graus de liberdade do totais do modelo.

2.5 Teste t para os coeficientes

Deseja-se verificar a significância individual dos coeficientes, para isso utiliza-se as hipóteses:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Tem-se que $\widehat{Var(\hat{\beta})} = QMResC_{jj}$. Em que C_{jj} é o j-ésimo elemento da diagonal da matriz $(\mathbf{X}'\mathbf{X})^{-1}$, QMRes é o quadrado médio do resíduo. Como $\hat{\beta}_j$ é uma combinação linear de distribuições normais, logo $\hat{\beta}_j$ segue uma distribuição Normal. Portanto, sob H_0, a estatística do teste é:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{QMResC_{jj}}} \sim t_{n-p-1} \tag{16}$$

com nível de significância α, a hipótese H_0 é rejeitada se $|t_0| > t_{(1-\frac{\alpha}{2},n-p-1)}$. Considerando o p-valor dado pela expressão

$$2P(t_{n-p-1} > |t_0|) \tag{17}$$

Rejeita-se H_0 se o p-valor $< \alpha$.

2.6 Teste F - ANOVA

$$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \\ H_1 : \text{Existe pelo menos um } \beta_k \neq 0, k = 0, 1, 2, ...p \end{cases}$$

A hipotése nula pode ser entendida como ausência de regressão e a hipótese alternativa como existência de regressão.

Tabela 1: Tabela de Análise de Variância

Fonte de Variação	**SQ**	**g.l**	**QM**	**F**
Regressão	SQReg	p	QMReg	QMReg/QMRes
Resíduo	SQRes	n - (p+1)	QMRes	
Total	SQTotal	n-1		

Sob H_0, a estatística do teste é:

$$F_0 = QMReg/QMRes \tag{18}$$

com distribuição F com p graus de liberdade no numerador e (n-(p+1)) graus de liberdade no denominador.

Considerando o p-valor, por meio da expressão:

$$P(F_{p,n-(p+1)} > |F_0|) \tag{19}$$

Rejeita-se H_0, se p-valor < α.

2.7 Transformação na variável resposta

Quando as suposições de homocedasticidade e normalidade do erro do modelo são violadas, é indicado realizar uma transformação na variável resposta para que a distribuição do resíduo se aproxime da distribuição normal. As transformações mais comuns são:

- $Y' = \sqrt{Y}$;
- $Y' = \ln Y$;
- $Y' = 1/Y$.

Neste estudo a transformação realizada foi a de $Y' = \ln Y$, de acordo com (Neter, Kutner, Nachtsheim, & Wasserman, 1996) usa-se essa transformação quando o erro padrão, para cada nível do resíduo, é proporcional a média do nível, ou seja, se σ_i for proporcional a μ_i a transformação indicada é $Y' = \ln Y$.

2.8 Seleção de modelos

Para a seleção de modelos foram usados os coeficientes de determinação, coeficientes de determinação ajustado, critérios de Mallows, o de informação Bayesiano, além disso também foram usados as seleções *Forward, Backward e Stepwise*.

2.8.1 Critério C_{p+1} de Mallows

O Critério C_{p+1} de Mallows é definido por:

$$C_{p+1} = \frac{SQRes_{p+1}}{MSRes} - (n - 2p - 2) \tag{20}$$

em que

- $SQRes_{p+1}$ é a soma de quadrados do resíduo do modelo de regressão ajustado com as p variáveis explicativas;
- $MSRes$ é o quadrado médio do resíduo com todas as variáveis explicativas.

2.8.2 Critério de informação Bayesiano - BIC_{p+1}

O critério de informação Bayesiano é definido por:

$$BIC_{p+1} = n\,\ln(SQRes_{p+1}) - n\,\ln(n) + (p+1)\,\ln(n) \tag{21}$$

Escolhe-se o modelo que apresenta o menor valor de BIC_{p+1} dentre todos os modelos considerados para o determinado problema.

2.8.3 Seleção Forward

Passo 1: Ajusta-se modelos de regressão simples com cada variável explicativa. Por meio da estatística t e seu respectivo p-valor, seleciona-se a variável mais significativa.

Passo 2: Considerando a variável selecionada no Passo 1, ajusta-se todos os possíveis modelos de regressão com duas variáveis, sendo a variável selecionada no Passo 1 presente em todos os modelos. Para cada modelo de regressão, calcula-se a estatística t para verificar o efeito de introduzir a variável X_k no modelo que já tem a variável selecionada no Passo 1. Se alguma variável for selecionada o processo continua. Caso contrário, o processo termina.

Passo 3: Supõe-se que alguma variável foi selecionada no Passo 2. Em seguida, ajusta-se todos os possíveis modelos com três variáveis, sendo as variáveis selecionadas nos Passos 1 e 2 presentes nos modelos. Para cada modelo de regressão, calcula-se a estatística t para verificar o efeito de introduzir a variável X_k no modelo. Se alguma variável for selecionada o processo continua. Caso contrário, o processo termina.

Passo 4: Repete-se o procedimento até o algoritmo não selecionar mais variáveis que devem entrar no modelo.

2.8.4 Eliminação *Backward*

Passo 1: Neste procedimento, primeiramente ajusta-se um modelo com todas as variáveis explicativas e calcula-se a estatística t e seu respectivo p-valor para cada variável do modelo. Se algum p-valor for maior que α estabelecido, então a variável é removida do modelo. Caso contrário, termina-se o procedimento.

Passo 2: Se foi removida alguma variável, ajusta-se um modelo com todas as variáveis que ficaram no modelo e calcula-se a estatística t e seu respectivo p-valor para cada variável do modelo. Se algum p-valor for maior que α estabelecido, então a variável é removida do modelo. Caso contrário, termina-se o procedimento.

Passo 3 Repete-se o procedimento até o algoritmo não identificar mais variáveis que devem ser retiradas do modelo.

2.8.5 Regressão *Stepwise*

Passo 1: Ajusta-se modelos de regressão simples com cada variável explicativa. Por meio da estatística t e seu respectivo p-valor, seleciona-se a variável mais significativa.

Passo 2: Considera-se a variável selecionada no Passo 1 presente em todos os modelos. Para cada modelo de regressão, calcula-se a estatística t para verificar o efeito de introduzir a variável X_k no modelo que já tem a variável selecionada no Passo 1. Se alguma variável for selecionada o processo continua. Caso contrário, o processo termina.

Passo 3: Supõe-se que alguma variável foi selecionada no Passo 2. Nesse passo, será verificado se a variável deve permanecer no modelo formado pelas variáveis selecionadas nos Passos 1 e 2. E por meio da estatística t, verifica-se se a variável selecionada no Passo 1 fica ou sai do modelo.

Passo 4: Supõe-se que as variáveis selecionadas nos Passos 1 e 2 ficam no modelo. O próximo passo é verificar qual é a próxima variável que deverá entrar no modelo e, em seguida, verifica-se qual(is) variável(is) que já estavam no modelo ficará(ão) ou sairá(ão). Esses passos vão se repetindo até chegar no modelo final.

2.9 Diagnóstico de Multicolinearidade

Neste estudo, foi verificado uma alta correlação entre algumas variáveis explicativas, o que é um indício que possa existir multicolinearidade no modelo, para verificar o efeito dessa multicolinearidade será usado o **Fator de inflação da variância (VIF)**.

O VIF para o coeficiente β_k é igual a:

$$(VIF)_k = (1 - R_k^2)^{-1} \tag{22}$$

Em que R_k^2 é o coeficiente de determinação múltiplo quanto X_k é regredido nas outras $p-1$ variáveis explicativas do modelo. O maior valor de VIF dentre todas as variáveis normalmente é usado como o indicador de multicolinearidade. Se o maior valor de VIF for superior a 10 então pode-se dizer que a multicolinearidade está influenciando as estimativas dos coeficientes.

2.10 Teste de Breusch-Pagan

Um teste para homocedasticidade de variância é o Breusch-Pagan, em que é necessário a suposição de Normalidade e independência do erro, e que a variância do termo $\varepsilon_i,$ denotada por σ_i^2 é relacionada aos níveis das variáveis X do seguinte modo:

$$\ln \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \cdots + \gamma_p X_{ip} \tag{23}$$

em que as hipóteses do teste são:

$$\begin{cases} H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_p \\ H_1 : \exists \gamma_k \neq 0, \quad k = 1, \ldots, p \end{cases}$$

Em que é efetuada a regressão do resíduo e_i^2 em relação às variáveis explicativas e obtém-se a soma de quadrados dessa regressão, denotada por $SQRes^*$ e a estatística do teste é χ^2_{BP} denotada por:

$$\chi^2_{BP} = \frac{SQRes^*}{p+1} \div \left(\frac{SQRes}{n}\right)^2 \tag{24}$$

$SQRes$ é a soma de quadrados de resíduos da regressão de Y em relação a $X_1, \ldots, X_p$, n é o tamanho da amostra que está sendo feita a regressão e χ^2_{BP} segue uma distribuição χ^2_p.

2.11 Teste de Normalidade de Shapiro-Wilk

O teste de Shapiro-Wilk para Normalidade possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição normal} \\ H_1 : \text{A variável não segue uma distribuição normal} \end{cases}$$

O teste é baseado na estatística W denotada por:

$$W = \frac{b^2}{\sum_{i=1}^{n} (X_{(i)} - \bar{X})^2} \tag{25}$$

Em que X_i são os valores da amostra ordenados, e b é constante denotada por:

$$b = \begin{cases} \sum_{i=1}^{\frac{n}{2}} a_{n-i+1} \times (X_{(n-i+1)} - X_{(i)}) \text{ se n é par} \\ \sum_{i=1}^{\frac{(n+1)}{2}} a_{n-i+1} \times (X_{(n-i+1)} - X_{(i)} \text{ se n é ímpar} \end{cases} \tag{26}$$

Em que a_{n-i+1} são constantes geradas pelas médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n de uma distribuição Normal. A estatística do teste é a estatística W definida acima.

2.12 Observações Discrepantes

Pontos potencialmente distantes tem impacto nas estimativas dos parâmetros, erro padrão, valores preditos e estatísticas do modelo. A matriz H

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

é importante para detectar observações influentes. Os elementos h_{ii} da matriz H são definidos por:

$$h_{ii} = \mathbf{X_i'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X_i}$$

,

Em que X_i é a i-ésima linha da matriz X. A diagonal da matriz H é uma medida padronizada da distância da i-ésima observação do centro do espaço de X.

Valor grande de h_{ii} indica que a i-ésima observação está distante do centro das observações e que essa observação pode ser considerada um ponto de alavanca.

2.13 Observações Influentes

2.13.1 DFFIT

Medida da influência que a i-ésima observação tem sobre o valor ajustado $\hat{Y}_i$ é dada por:

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSRes_{(i)} h_{ii}}} \tag{27}$$

Em que $\hat{Y}_i$ é o valor ajustado para o i-ésimo caso quando todas as n observações são utilizadas no ajuste do modelo. $\hat{Y_{i(i)}}$ é o valor ajustado para o i-ésimo caso quando o i-ésimo caso é omitido no ajuste do modelo. $MSRes_{(i)}$ é o $MSRes$ calculado quando o i-ésimo caso é omitido no ajuste do modelo.

Para identificar se uma observação é influente, pode-se usar os seguintes critérios:

- $|(DFFITS)_i| > 1$ para conjuntos de dados pequenos ou médios indica que a observação é influente.
- $|(DFFITS)_i| > 2\sqrt{(p+1)/n}$ para conjuntos de dados grandes indica que a observação é influente.

2.13.2 Distância de Cook

Essa medida tem o intuito de verificar o efeito de excluir a i-ésima observação sobre o coeficiente β. Esta medida é definida como:

$$D_i = \sum_{i=1}^{n} \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{(p+1)MSRes} \tag{28}$$

Observações com $D_i > 1$ são consideradas observações influentes.

2.13.3 DFBETA

A medida da influência que a i-ésima observação tem sobre cada um dos coeficientes de regressão estimados é dada por:

$$(DFBETA)_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSRes_i C_{kk}}} \tag{29}$$

Em que β_k é o coeficiente de regressão estimado considerando todos os n casos no ajuste do modelo. $\hat{\beta}_{k(i)}$ é o coeficiente de regressão estimado considerando que o i-ésimo caso é omitido no ajuste do modelo. $MSRes_i$ é o $MSRes$ quando o i-ésimo caso é omitido no ajuste do modelo. C_{kk} é o k-ésimo elemento da diagonal principal da matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

Interpretação das medidas:

- Sinal - indica se a inclusão do i-ésimo caso leva ao aumento ou à diminuição do coeficiente de regressão estimado.
- Valor grande de $(DFBETA)_{k(i)}$ indica grande influência do i-ésimo caso na estimativa do k-ésimo coeficiente de regressão.
- $|(DFBETA)_{k(i)}| > 1$ para conjuntos de dados pequenos ou médios implica que a observação é influente.
- $|(DFBETA)_{k(i)}| > 2/\sqrt{n}$ para conjuntos de dados grandes implica que a observação é influente.

3 Resultados - Amostra de Desenvolvimento

Nesta seção, serão apresentados os resultados que foram encontrados ao aplicar os métodos elencados na metodologia ao conjunto de dados utilizado no estudo como amostra de desenvolvimento.

3.1 Análise Descritiva

Serão realizados estudos exploratórios com o objetivo de identificar possíveis comportamentos individuais das variáveis que sejam distantes do esperado e identificar as variáveis que provavelmente influenciam mais o valor da variável resposta, que é o consumo do carro em milhas por galão (*Mpg*).

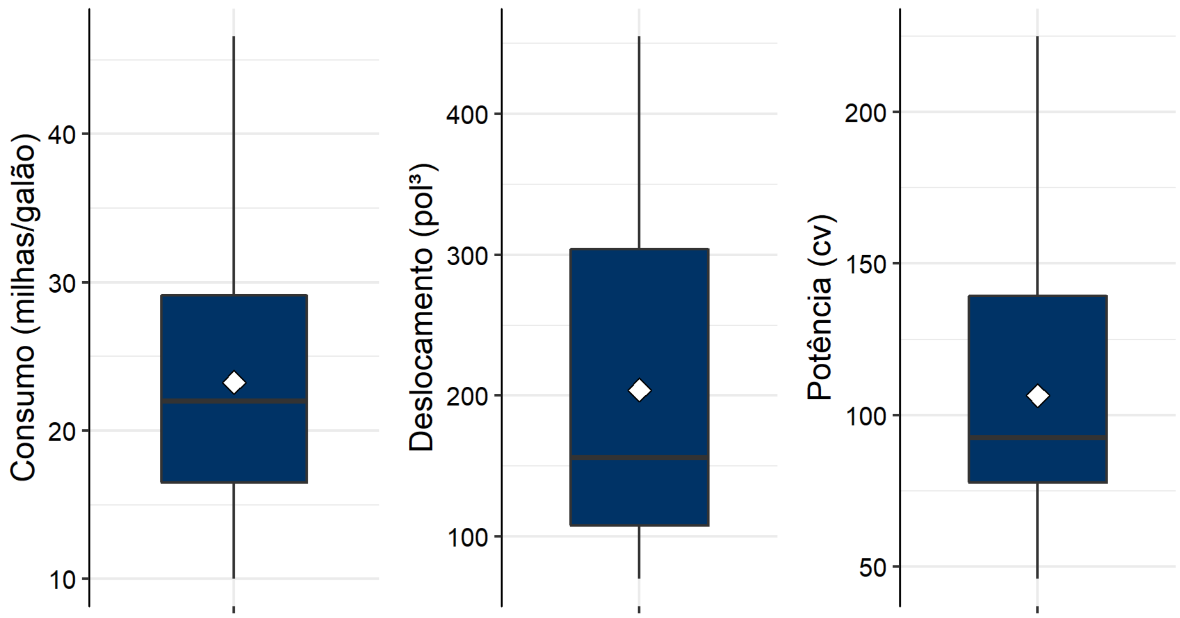
3.1.1 Univariada

Para estudar o comportamento individual das variáveis, segue abaixo um quadro com as medidas-resumo de cada variável quantitativa e um painel com gráficos ilustrando o comportamento de cada variável.

Quadro 1: Medidas-resumo de cada variável quantitativa

Variável	Mín	1º Quartil	Mediana	Média	3º Quartil	Máx	Desvio padrão
Mpg	10	16,5	22	23,25	29,12	46,6	7,92
Cylinders	3	4	6	5,59	8	8	1,73
Displacement	70	108	156	204,05	304	455	106,6
Horsepower	46	77,75	92,5	106,46	139,25	225	38,23
Weight	1613	2285,75	2912	3053,27	3756,5	5140	867,96
Acceleration	9	13,5	15,5	15,38	17	24,6	2,59
Age	1	3,75	7	6,8	10	13	3,71

Figura 3: Painel das variáveis



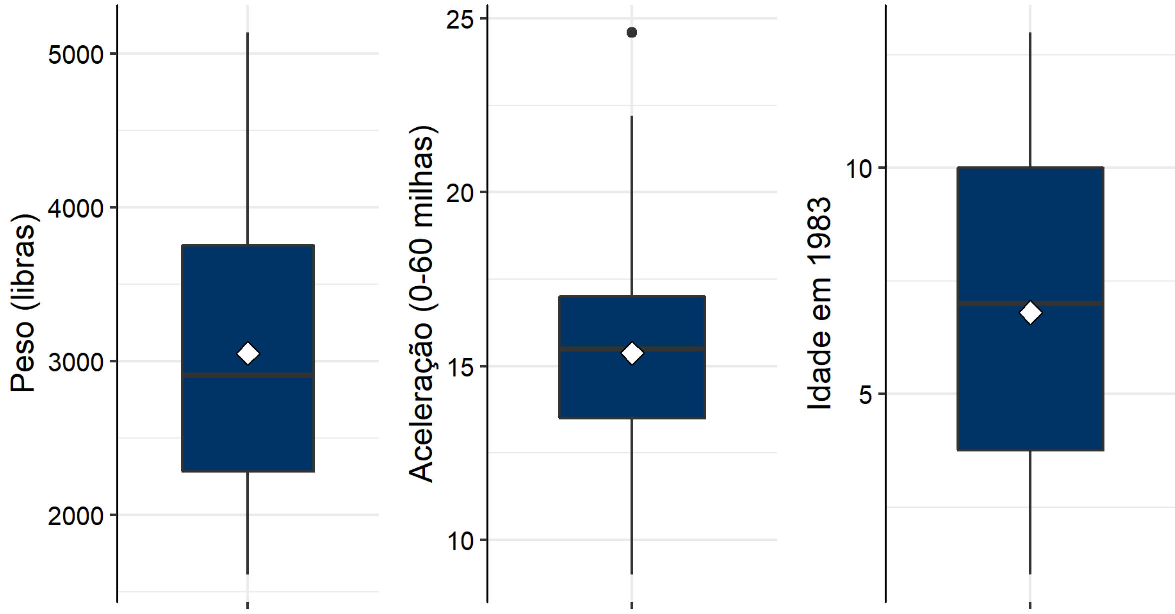
Considerando a Figura 3, pode-se analisar os gráficos *Boxplot* para as variáveis *Mpg*, *Displacement* e *Horsepower*.

Para a variável *Mpg*, pode-se verificar que não há valores extremos e que a média é um pouco acima da mediana, como pode-se comprovar observando o Quadro 1 em que a média é de 23,25 enquanto a mediana é de 22.

Enquanto para a variável *Displacement* é possível identificar que a média é muito superior à mediana, o que pode-se comprovar pelo Quadro 1 em que a média é de 204,05 e a mediana é de 156. Também é possível verificar que a média está acima de 50% dos valores, logo, existe uma forte assimetria à direita na forma em que os valores estão distribuídos.

A variável *Horsepower* também possui a média acima da mediana, já que a dispersão dos valores acima da mediana é maior que a de valores abaixo da mesma. Também é possível verificar que, observando o Quadro 1, a média é igual a 106,46 enquanto a mediana é de 92,5.

Figura 4: Painel das variáveis



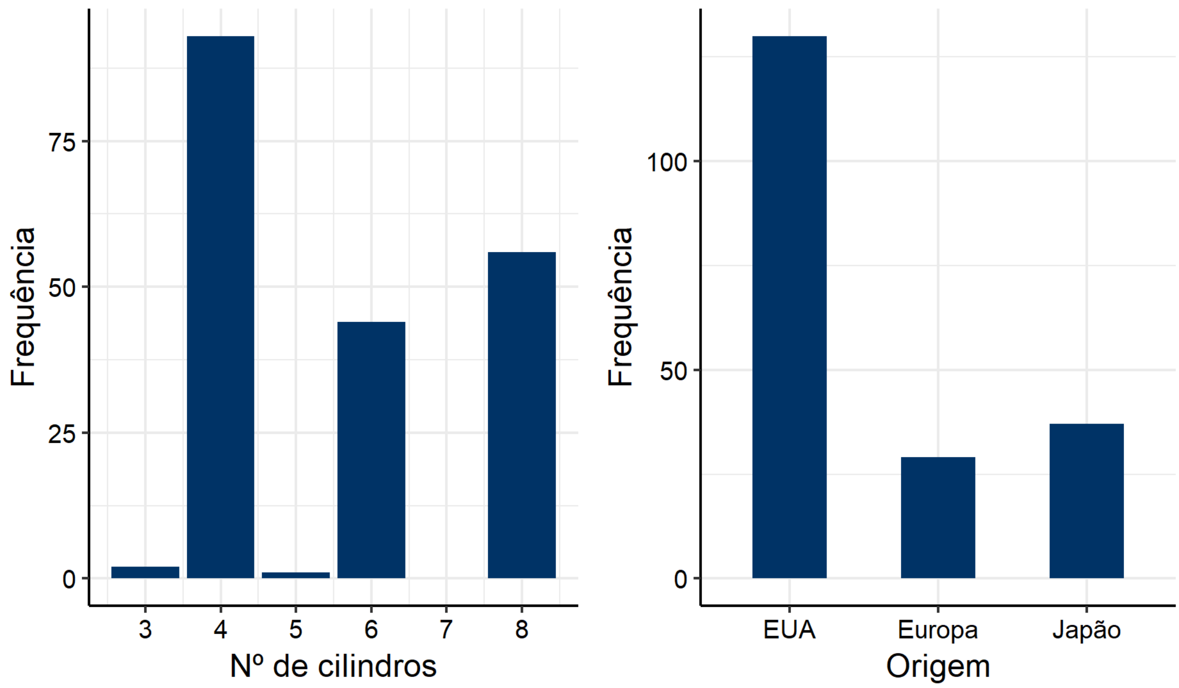
Considerando a Figura 4, pode-se analisar os gráficos *Boxplot* para as variáveis *Weight*, *Acceleraton* e *Age*.

A variável *Weight* não apresenta valores extremos e a média é um pouco acima da mediana, como pode-se comprovar observando o Quadro 1, em que a média é de 3053,27, enquanto a mediana é de 2912. Também é possível verificar que a amplitude da variável é de 1613 libras até 5140.

Enquanto para a variável *Acceleration*, é possível identificar que há um valor extremo superior, ou seja, um valor que ultrapassa o limite superior do *Boxplot*. Também é possível notar que a média é muito próxima da mediana, o que pode-se comprovar de acordo com o Quadro 1, em que a média é de 15,38 enquanto a mediana é de 15,5. É possível ver que os dados parecem estar igualmente dispersos.

A variável *Age* também possui a média muito próxima da mediana. Além disso, é possível verificar que a média é igual a 6,8 enquanto a mediana é de 7.

Figura 5: Painel das variáveis



Considerando a Figura 5, pode-se analisar as variáveis *Cylinders* e também *Origin*.

Observando o gráfico para *Cylinders*, pode-se notar que o valor predominante é de 4 cilindros nos motores dos carros. Não há nenhuma observação com 7 cilindros. Também é possível notar que os dados estão distribuídos de maneira heterogênea, visto que os valores de 4, 6 e 8 cilindros concentram a grande maioria das observações. De acordo com o Quadro 1, pode-se verificar que a mediana é igual a 6, ou seja, 50% dos valores estão abaixo de 6.

Analisando a variável *Origin*, pode-se verificar que a origem com maior frequência é EUA. Além disso, é possível verificar que o Japão é a segunda origem mais frequente, porém ainda é muito inferior aos EUA. E, por fim, a origem que teve menor frequência absoluta foi Europa.

3.1.2 Bivariada

Com o intuito de estudar o comportamento bivariado da variável resposta (*Mpg*) com cada variável explicativa, elaborou-se um painel ilustrando esse comportamento e também um quadro contendo os coeficientes de correlação de Pearson. Para o caso em que ambas as variáveis eram qualitativas, utilizou-se o coeficiente de contingência modificado. Para os casos em que uma variável era qualitativa e a outra era quantitativa, utilizou-se o coeficiente R^2.

Quadro 2: Matriz de Correlação

Correlação	*Mpg*	*Cylinders*	*Displacement*	*Horsepower*	*Weight*	*Acceleration*	*Age*	*Origin*
Mpg	1	-0,788	-0,806	-0,772	-0,836	0,413	-0,62	0,337
Cylinders	-0,788	1	0,954	0,853	0,91	-0,486	0,392	0,622
Displacement	-0,806	0,954	1	0,899	0,929	-0,536	0,412	0,683
Horsepower	-0,772	0,853	0,899	1	0,859	-0,704	0,456	0,483
Weight	-0,836	0,91	0,929	0,859	1	-0,425	0,37	0,543
Acceleration	0,413	-0,486	-0,536	-0,704	-0,425	1	-0,309	0,103
Age	-0,62	0,392	0,412	0,456	0,37	-0,309	1	0,025
Origin	0,337	0,622	0,683	0,483	0,543	0,103	0,025	1

Figura 6: Painel das variáveis explicativas versus a variável resposta (*Mpg*)

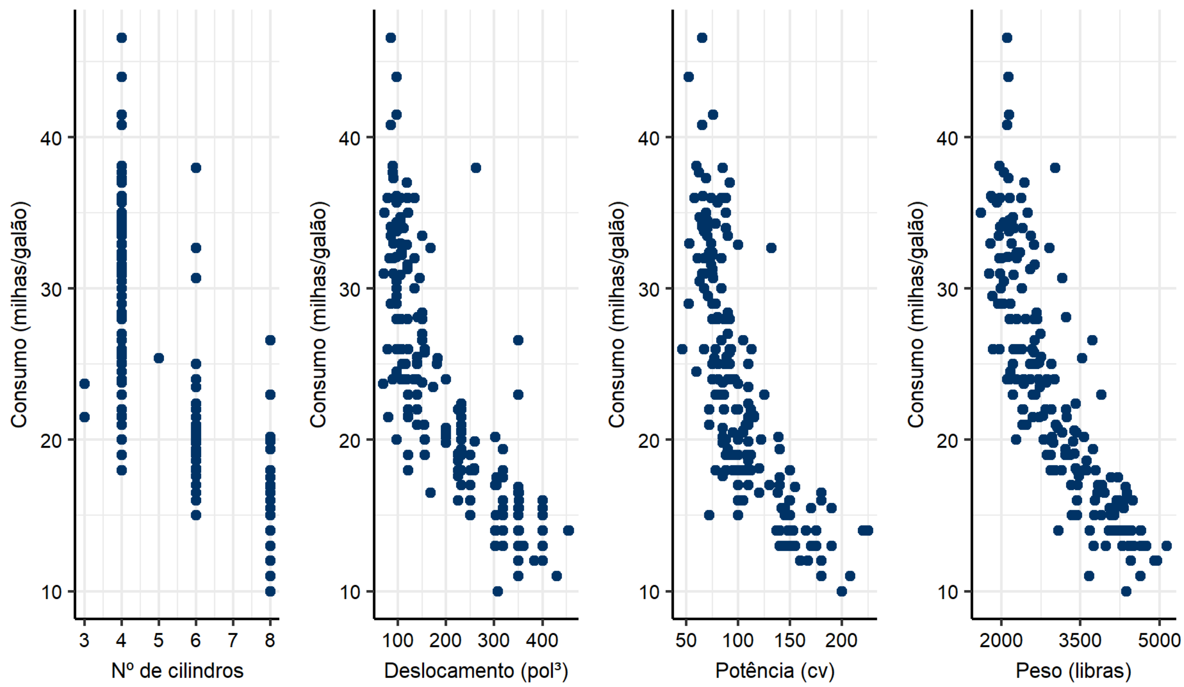
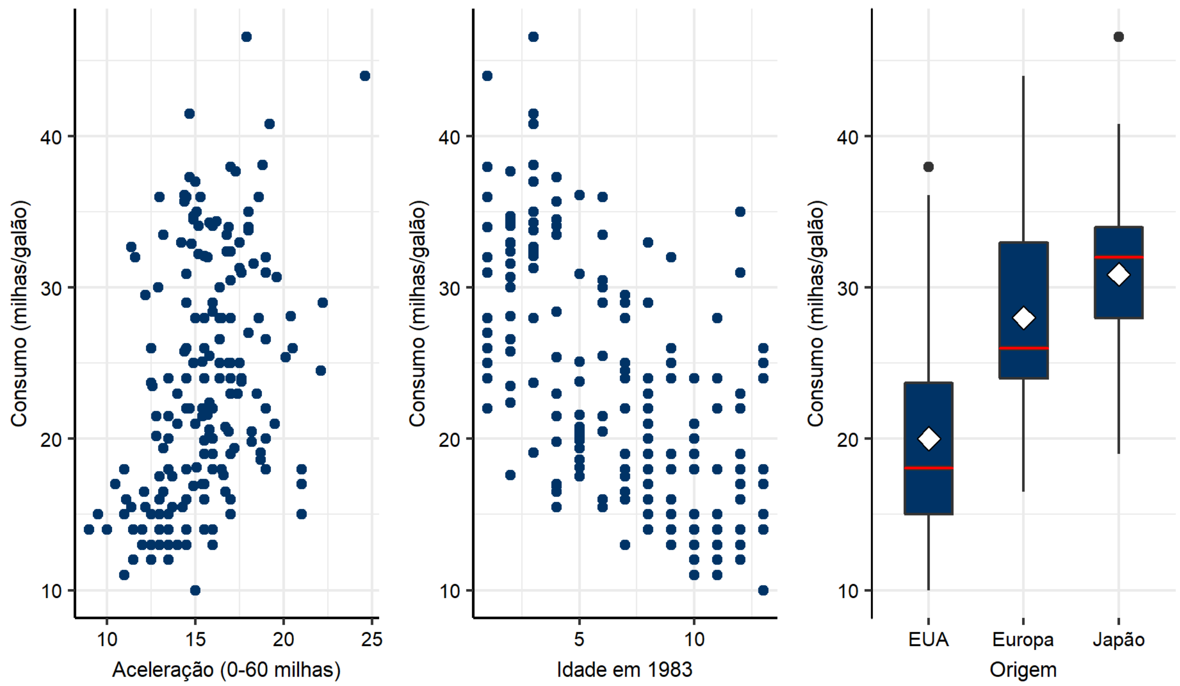


Figura 7: Painel das variáveis explicativas versus a variável resposta (*Mpg*)



Ao observar a Figura 6, verifica-se que provavelmente a relação entre o número de cilindros do motor do carro e seu consumo é inversamente proporcional, o que é corroborado pelo valor do coeficiente de correlação de Pearson mostrado no Quadro 2, que é de -0,788, valor indicativo de correlação forte negativa entre as variáveis.

Para as variáveis Deslocamento do motor, Potência e Peso do carro, os gráficos de dispersão da Figura 6 ilustram um comportamento conjunto com a variável consumo do carro semelhante: as três aparentam ter correlação negativa forte com a variável consumo. A matriz de Correlação no Quadro 2 ratifica isto, pois os coeficientes de correlação de Pearson entre a variável resposta consumo do carro (*Mpg*) e as variáveis explicativas Deslocamento (*Displacement*), Potência (*Horsepower*) e Peso (*Weight*) são, respectivamente, -0,806, -0,772 e -0,836.

Quanto à variável Aceleração (0-60 milhas), percebe-se o oposto do observado para as variáveis analisadas acima: o gráfico indica uma possível correlação fraca ou moderada positiva entre essa variável e o consumo do carro, o que é apontado pela Matriz de Correlação do Quadro 2, pois o coeficiente de correlação de Pearson entre essa variável e *Mpg* é 0,413.

Já para a variável Idade em 1983, observou-se comportamento semelhante às outras variáveis com exceção de Aceleração. Isto é, há indicativos de correlação negativa tanto no gráfico de dispersão da Figura 7 quanto na Matriz de Correlação do Quadro 2, pois o coeficiente de correlação de Pearson calculado foi de -0,62.

Por último, ao analisar a variável Origem, que é a única variável qualitativa, verifica-se que: os carros com origem Japão aparentam apresentar maior consumo do que os carros Europeus, que, por sua vez, parecem possuir maior consumo que os carros Americanos, o que é ratificado pelo terceiro gráfico do painel da Figura 7.

3.2 Modelo Completo

De início, ajustou-se o modelo completo contendo todas as variáveis do banco de dados. A variável *Origin* foi dividida em duas variáveis indicadores (*Japan* e *Europe*), pois a variável *Origin* possui 3 categorias. A equação do modelo é:

$$\begin{aligned} Mpg_i = \beta_0 + \beta_1 Cylinders_i + \beta_2 Displacement_i + \beta_3 Horsepower_i + \beta_4 Weight_i \\ + \beta_5 Acceleration_i + \beta_6 Age_i + \beta_7 Japan_i + \beta_8 Europe_i + \varepsilon_i \end{aligned} \tag{30}$$

Abaixo, seguem os principais resultados para esse modelo.

Quadro 3: Resultados dos coeficientes

Variável	Estimativa	Erro padrão	Estatística t	P-valor
Intercept	47,9127	3,1419	15,2497	<0,0001
Cylinders	-0,4469	0,4755	-0,9399	0,3485
Displacement	0,0206	0,0105	1,9566	0,0519
Horsepower	-0,0343	0,0206	-1,6641	0,0978
Weight	-0,0057	9e-04	-6,6578	<0,0001
Acceleration	-0,0828	0,1502	-0,5513	0,5821
Age	-0,7698	0,0731	-10,5287	<0,0001
Japan	3,6721	0,8116	4,5246	<0,0001
Europe	2,9054	0,8433	3,4453	0,0007

Observando o Quadro 3, pode-se notar a significância de cada variável no modelo completo. Sendo assim, considerando o nível de significância $\alpha = 0,05$, nota-se que as variáveis *Cylinders*, *Displacement*, *Horsepower* e *Acceleration* não são significativas, pois apresentam p-valores respectivamente iguais a: 0,3485; 0,0519; 0,0978; 0,5821.

Quadro 4: Resultados do resíduo

Erro padrão do resíduo	G.L.
3,3276	187

De acordo com o Quadro 4, observa-se que o erro padrão estimado do modelo, ou seja, a estimativa de σ possui o valor de 3,3276 e 187 graus de liberdade.

Quadro 5: Coeficientes de Determinação

$\mathbf{R^2}$	$\mathbf{R^2_{adj}}$
0,8307	0,8235

Observando o Quadro 5, tem-se que R^2 e R^2_{afj} são respectivamente iguais a 0,8307 e 0,8235. Sendo assim, pode-se concluir que o modelo completo é capaz de explicar 83,07% da variável resposta para o caso do R^2. Também pode-se dizer que, ajustado ao número de vaiáveis explicativas, o modelo completo é capaz de explicar 82,35% da variável resposta.

Quadro 6: Resultados principais da Análise de Variância

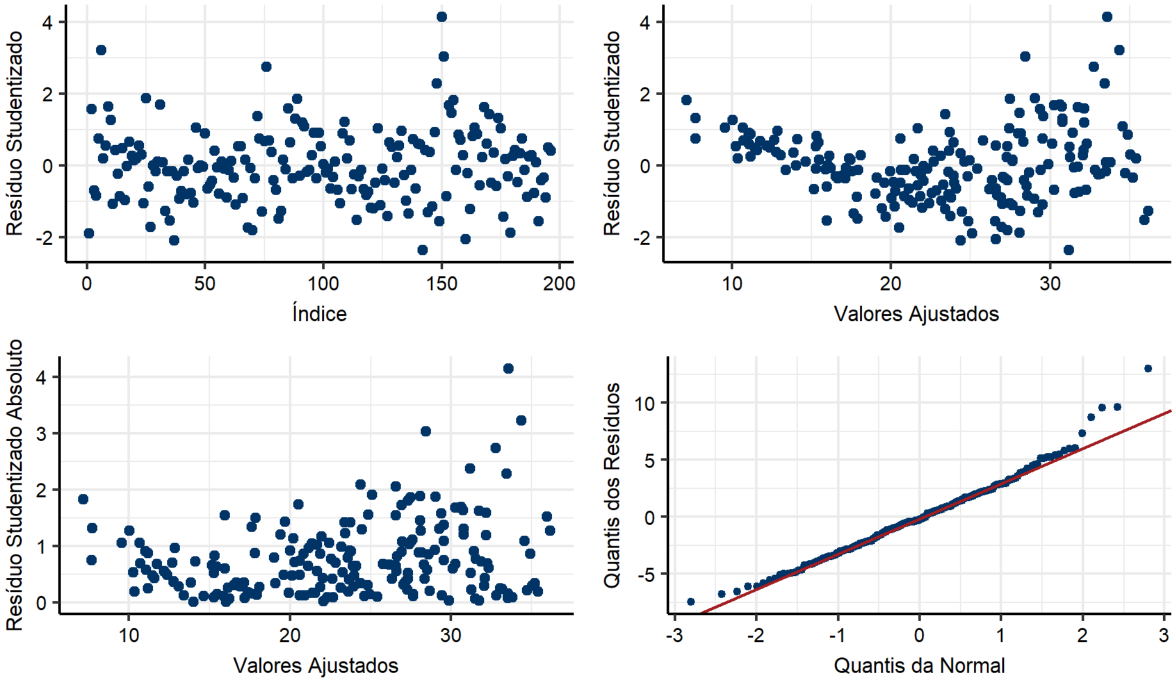
Estatística F	**G.L. numerador**	**G.L. denominador**	**P-valor**
114,7129	8	187	<0,0001

Observando o Quadro 6, considerando o nível de confiança $\alpha = 0,05$, pode-se concluir que há regressão.

3.2.1 Verificação dos Pressupostos

Para verificação dos pressupostos, foram utilizados métodos gráficos e testes de hipótese.

Figura 8: Painel de verificação dos pressupostos



Analisando a Figura 8, pode-se verificar os pressupostos do modelo completo. O gráfico de dispersão dos resíduos *studentizados* pelo índice indica a independência das observações, pois apresenta aleatoriedade aparente.

Os gráficos referentes aos valores ajustados pelos resíduos *studentizados* fornecem informação a respeito da homocedasticidade e linearidade do modelo. Nesse caso, pode-se dizer que o gráfico não está de acordo com o pressuposto, pois o gráfico ilustra que os resíduos *studentizados* aparentam ser menos dispersos para valores ajustados menores, e mais dispersos para valores ajustados maiores.

O gráfico de probabilidade normal está claramente fora do pressuposto, visto que há um desvio da linha representando o esperado sob normalidade. Ou seja, pode-se concluir que os resíduos não seguem uma distribuição normal.

Quadro 7: Testes de Hipótese dos pressupostos

Teste	**Estatística do Teste**	**P-valor**
Breusch-Pagan	6,486	0,593
Shapiro-Wilk	0,9812	0,0098

O Quadro 7 possui as informações referentes aos testes de hipótese para homocedasticidade e normalidade dos resíduos. Considerando o nível de significância de $\alpha = 0,05$, pode-se concluir que a homocedasticidade do modelo é verificada pelo teste de *Breusch-Pagan* com p-valor igual a 0,593. Porém, o pressuposto de normalidade não é verificado pelo teste de Shapiro-Wilk, visto que o p-valor de 0,0098 é menor que o nível de significância.

3.3 Modelo Transformado

Verificou-se a ausência de normalidade dos resíduos. Por isso, foi utilizada uma transformação na variável resposta *Mpg*. A transformação escolhida foi $Mpg' = \ln{(Mpg)}$. A equação do modelo é:

$$\ln{(Mpg)_i} = \beta_0 + \beta_1 Cylinders_i + \beta_2 Displacement_i + \beta_3 Horsepower_i + \beta_4 Weight_i \\ + \beta_5 Acceleration_i + \beta_6 Age_i + \beta_7 Japan_i + \beta_8 Europe_i + \varepsilon_i \quad (31)$$

Abaixo, seguem os principais resultados para esse modelo.

Quadro 8: Resultados dos coeficientes

Variável	Estimativa	Erro padrão	Estatística t	P-valor
Intercept	4,2602	0,111	38,3682	<0,0001
Cylinders	-0,0235	0,0168	-1,397	0,1641
Displacement	0,0007	0,0004	1,7908	0,0749
Horsepower	-0,002	0,0007	-2,7703	0,0062
Weight	-0,0002	$3 \cdot 10^{-5}$	-7,5856	<0,0001
Acceleration	-0,0057	0,0053	-1,0821	0,2806
Age	-0,0312	0,0026	-12,0579	<0,0001
Japan	0,1149	0,0287	4,0047	0,0001
Europe	0,097	0,0298	3,2541	0,0013

De acordo com Quadro 8 e considerando o nível de significância $\alpha = 0,05$, pode-se concluir que, para o modelo transformado, as variáveis *Cylinders*, *Displacement* e *Acceleration* apresentam p-valores respectivamente iguais a: 0,1661; 0,074 e 0,2806.

É importante ressaltar o fato de que *Horsepower* é significativa para o modelo transformado, sendo que não era para o modelo completo.

Quadro 9: Resultados do resíduo

Erro padrão do resíduo	G.L.
0,1176	187

De acordo com o Quadro 9, observa-se que o erro padrão estimado do modelo, ou seja, a estimativa de σ possui o valor de 0,1176 e 187 graus de liberdade. Esse valor é bem diferente do encontrado no modelo completo original, visto que o erro padrão diminui bruscamente de um valor de 3,3276 para 0,1176.

Quadro 10: Coeficientes de Determinação

$\mathbf{R^2}$	$\mathbf{R^2_{adj}}$
0,8882	0,8834

Observando o Quadro 10, tem-se que R^2 e R^2_{adj} são respectivamente iguais a 0,8882 e 0,8834. Sendo assim, pode-se concluir que o modelo transformado é capaz de explicar 88,82% da variável resposta para o caso do R^2. Também pode-se dizer que, ajustado ao número de vaiáveis explicativas, o modelo transformado é capaz de explicar 88,34% da variação da variável resposta considerando R^2_{afj}.

Logo, o modelo transformado tem um poder de explicação da variabilidade da variável resposta maior que o modelo completo original.

Quadro 11: Resultados principais da Análise de Variância

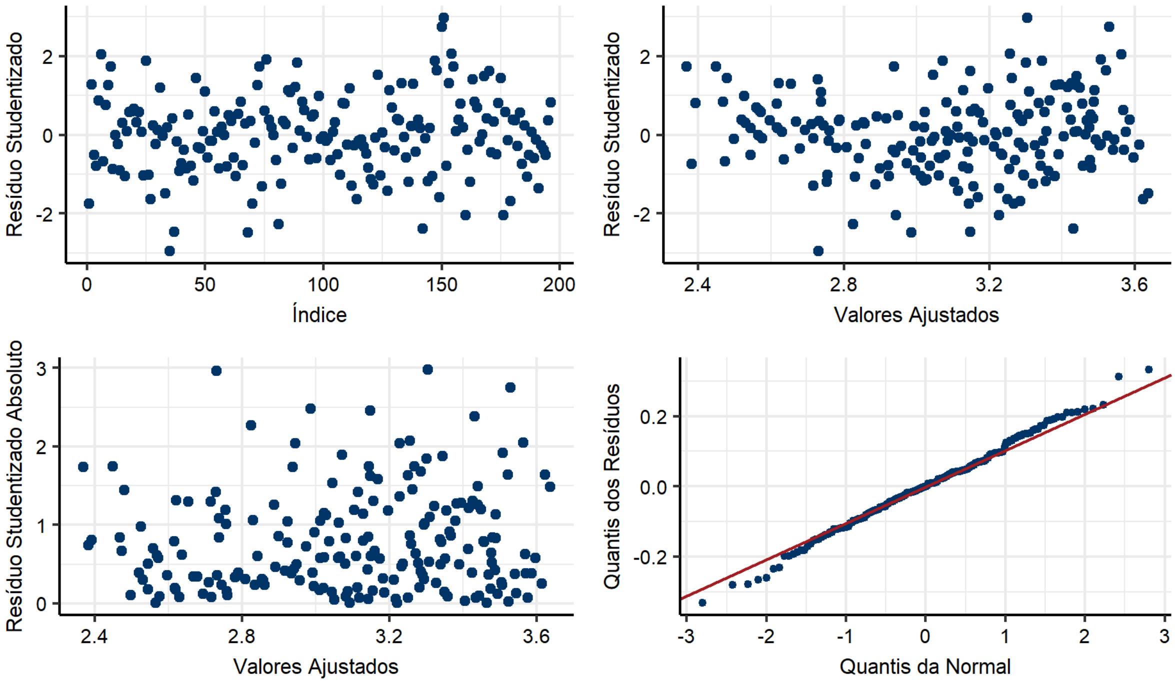
Estatística F	**G.L. numerador**	**G.L. denominador**	**P-valor**
185,751	8	187	<0,0001

Considerando o nível de significância $\alpha = 0,05$, analisando o Quadro 11, pode-se concluir que há regressão quando considera-se o modelo transformado.

3.3.1 Verificação dos Pressupostos

Para verificação dos pressupostos, foram utilizados métodos gráficos e testes de hipótese.

Figura 9: Painel de verificação dos pressupostos



Analisando a Figura 9, pode-se verificar os pressupostos do modelo completo transformado. O gráfico dos resíduos *studentizados* pelo índice indica que há independência das observações, pois esse gráfico apresenta aleatoriedade aparente.

Os gráficos referentes aos valores ajustados por resíduos *studentizados* fornecem informação a respeito da homocedasticidade e linearidade do modelo. Nesse caso, nota-se que os gráficos confirmam os dois pressupostos, visto que não parece haver nenhum tipo de padrão não aleatório nos resíduos *studentizados* a medida que o valor dos valores ajustados aumenta.

O gráfico de probabilidade normal está bem mais próximo do esperado sob normalidade do que o modelo completo original, visto que não há um desvio significativo da linha representando os quantis da distribuição normal. Ou seja, pode-se concluir que os resíduos seguem uma distribuição normal.

Quadro 12: Testes de Hipótese dos pressupostos

Teste	Estatística do Teste	P-valor
Breusch-Pagan	3,0551	0,9309
Shapiro-Wilk	0,9958	0,8678

O Quadro 7 possui as informações referentes aos testes de hipótese para homocedasticidade e normalidade dos resíduos. Considerando o nível de significância de

$\alpha = 0,05$, pode-se concluir que a homocedasticidade do modelo é verificada pelo teste de *Breusch-Pagan* com p-valor igual a 0,9309, que é muito superior ao p-valor encontrado para o modelo completo original.

Por sua vez, o pressuposto de normalidade é verificado pelo teste de Shapiro-Wilk, visto que o p-valor de 0,8676 é maior que o nível de significância. Dessa maneira, conclui-se que a transformação obteve sucesso em tornar os pressupostos válidos.

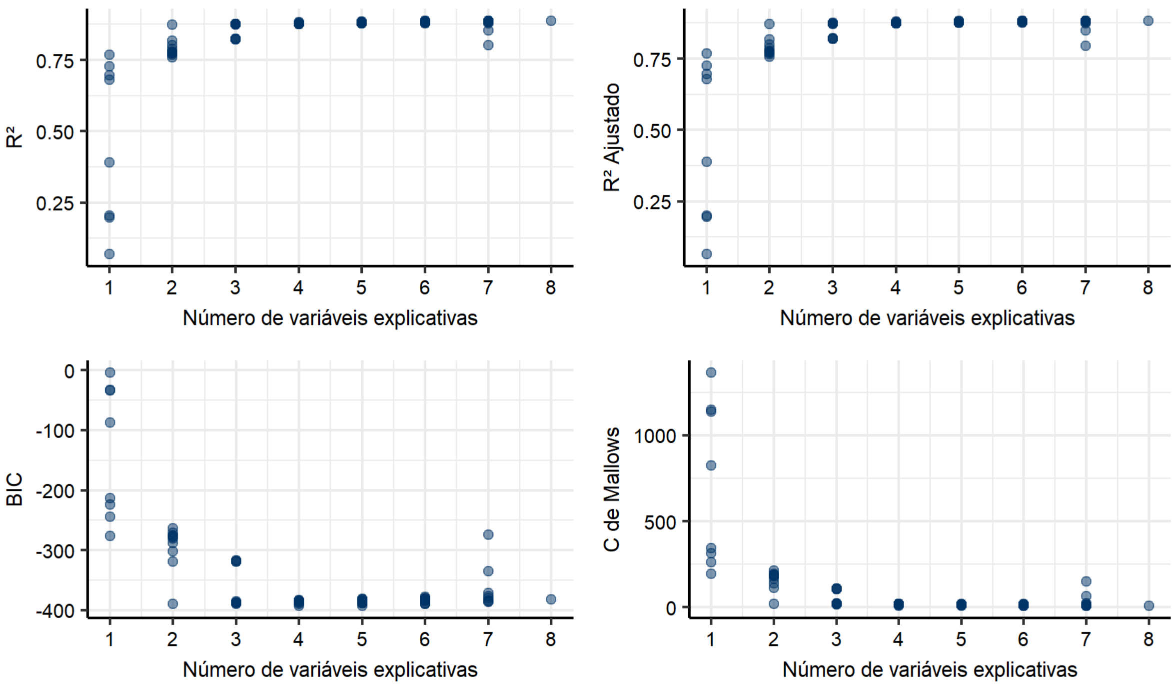
3.4 Seleção de Variáveis

Os modelos candidatos serão determinados utilizando três métodos: critérios de seleção, eliminação *backward* manual (utilizando significância do teste t), seleção automática utilizando AIC (eliminação *backward*, seleção *forward* e regressão *stepwise*).

3.4.1 Critérios de Seleção

Os critérios de seleção que serão considerados são: R^2, R_a^2dj, BIC, $C_{Mallows}$. Abaixo, seguem os gráficos do número de variáveis explicativas pelo valor de cada um dos critérios.

Figura 10: Painel de seleção de variáveis



Pela Figura 10, conclui-se que devem ser considerados modelos com 2 a 6 variáveis explicativas. Foram selecionados como modelos candidatos os melhores modelos para cada quantidade de variáveis explicativas, ou seja, tem-se um modelo candidato com 2, um com 3, e assim por diante. Para definir qual era o melhor modelo, utilizou-se os critérios de seleção. Os modelos candidatos são:

$$\text{Modelo 1: } \ln{(Mpg)}_i = \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \varepsilon_i \tag{32}$$

$$\text{Modelo 2: } \ln{(Mpg)}_i = \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Japan_i + \varepsilon_i \tag{33}$$

$$\text{Modelo 3: } \ln{(Mpg)}_i = \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Japan_i + \beta_4 Europe_i + \varepsilon_i \tag{34}$$

$$\begin{aligned}\text{Modelo 4: } \ln{(Mpg)}_i = &\ \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Japan_i + \beta_4 Europe_i \\ &+ \beta_5 Horsepower_i + \varepsilon_i\end{aligned} \tag{35}$$

$$\begin{aligned}\text{Modelo 5: } \ln{(Mpg)}_i = &\ \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Japan_i + \beta_4 Europe_i \\ &+ \beta_5 Horsepower_i + \beta_6 Displacement_i + \varepsilon_i\end{aligned} \tag{36}$$

3.4.2 Eliminação *Backward* Manual

Foi determinado um modelo candidato utilizando eliminação *backward* manual (utilizando significância do teste t). O modelo encontrado segue abaixo, que é igual ao modelo 4 encontrado pelos critérios de seleção.

$$\begin{aligned}\ln{(Mpg)}_i = &\ \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Japan_i + \beta_4 Europe_i \\ &+ \beta_5 Horsepower_i + \varepsilon_i\end{aligned} \tag{37}$$

3.4.3 Seleção Automática

Foram determinados modelos candidatos utilizando os métodos de seleção automática que utiliza o AIC como critério (eliminação *backward*, seleção *forward* e regressão *stepwise*). Os modelos encontrados seguem abaixo.

$$\textit{Backward: } \ln (Mpg)_i = \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Japan_i + \beta_4 Europe_i \\ + \beta_5 Horsepower_i + \beta_6 Displacement_i + \beta_7 Cylinders_i + \varepsilon_i \tag{38}$$

$$\textit{Forward: } \ln (Mpg)_i = \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Japan_i + \beta_4 Europe_i \\ + \beta_5 Horsepower_i + \varepsilon_i \tag{39}$$

$$\textit{Stepwise: } \ln (Mpg)_i = \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Japan_i + \beta_4 Europe_i \\ + \beta_5 Horsepower_i + \varepsilon_i \tag{40}$$

Os modelos encontrados pelo *forward* e *stepwise* são iguais ao modelo 4 encontrado pelos critérios de seleção. Logo, adiciona-se apenas um modelo candidato aos 4 já definidos.

$$\textit{Modelo 6: } \ln (Mpg)_i = \beta_0 + \beta_1 Age_i + \beta_2 Weight_i + \beta_3 Japan_i + \beta_4 Europe_i \\ + \beta_5 Horsepower_i + \beta_6 Displacement_i + \beta_7 Cylinders_i + \varepsilon_i \tag{41}$$

3.4.4 Seleção final

Abaixo, segue um quadro apresentando os valores encontrados para os critérios de seleção para cada modelo candidato.

Quadro 13: Critérios de Seleção dos Modelos Candidatos

Modelo	R^2	R^2_{adj}	BIC	$C_{Mallows}$
1	0,8737	0,8724	-389,6876	21,3232
2	0,8771	0,8752	-389,812	17,5779
3	0,8823	0,8798	-392,9755	10,9117
4	0,8855	0,8825	-393,1281	7,5307
5	0,8864	0,8828	-389,3154	8,104
6	0,8875	0,8833	-386,0404	8,171

Os modelos 5 e 6 apresentaram variáveis não significativas, utilizando o teste t. Dentre os modelos 1 a 4, foram selecionados para realizar a análise de diagnóstico os dois que apresentaram as melhores medidas no Quadro 13: Modelo 3 e Modelo 4.

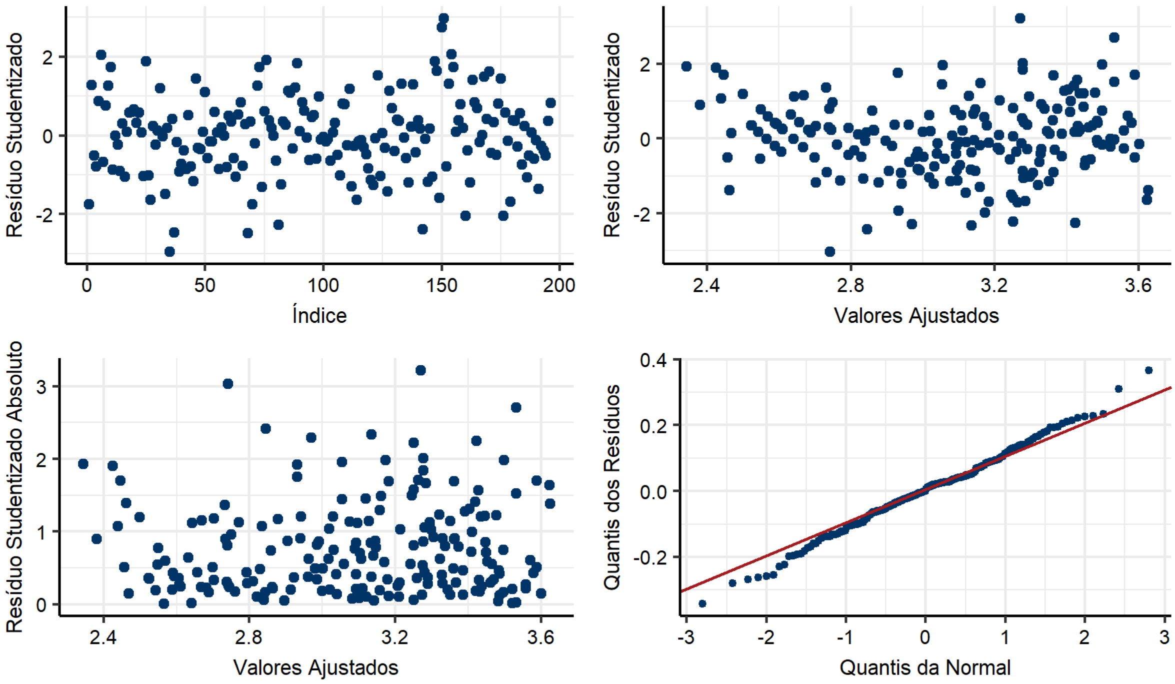
3.5 Modelo 4

Considerando que o Modelo 4 foi selecionado como modelo candidato, foi feita a análise de diagnóstico para verificar a adequação desse modelo.

3.5.1 Verificação de Pressupostos

Para a verificação dos pressupostos foram utilizados métodos gráficos e testes de hipótese.

Figura 11: Painel de verificação dos pressupostos



Considerando a Figura 11, pode-se verificar os pressupostos do modelo 4. O gráfico dos resíduos *studentizados* pelo índice verifica a independência das observações, já que aparenta ser aleatório.

Os gráficos referentes aos valores ajustados por resíduos *studentizados* servem para verificar a homocedasticidade e linearidade do modelo. Nesse caso, nota-se que os gráficos estão de acordo com o pressuposto, visto que não parece haver nenhum tipo de padrão não aleatório no comportamento dos resíduos.

O gráfico de probabilidade normal está bem ajustado, visto que não há um desvio significativo da linha representando os quantis da distribuição normal. Ou seja, é um indicio de que os resíduos do modelo 4 seguem uma distribuição normal.

Quadro 14: Testes de Hipótese dos pressupostos

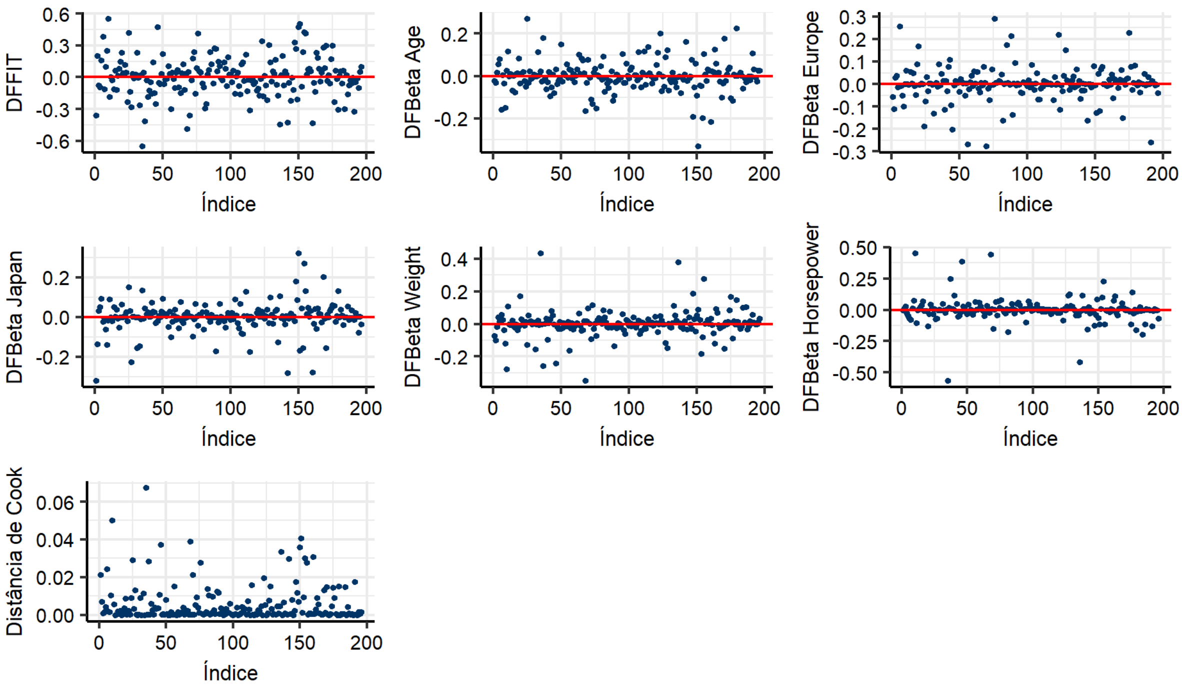
Teste	Estatística do Teste	P-valor
Breusch-Pagan	4,7117	0,452
Shapiro-Wilk	0,9952	0,7958

Considerando as informações do Quadro 14 e também o nível de significância de $\alpha = 0,05$, pode-se concluir que o modelo 4 atende ao pressuposto de homocedasticidade, visto que o teste de *Breusch-Pagan* possui p-valor de 0,452, maior que α. O pressuposto de normalidade também é atendido visto que o teste de Shapiro-Wilk tem p-valor de 0,7958, também maior que α.

3.5.2 Medidas influentes

Desejamos verificar se o modelo 4 possui alguma estimativa que é influenciada por valores extremos. Para isso, serão utilizadas as medidas $DFBeta$, $DFFIT$ e a Distância de Cook.

Figura 12: Medidas influentes do modelo 4



Depois de analisado os gráficos da Figura 12, obteve-se as seguintes observações influentes para o modelo 4: 10, 35, 68, 136. Quando retiradas do modelo 4, alteraram de maneira considerável a significância da variável *Horsepower*, também alterando significativamente o valor do estimador do coeficiente de *Horsepower*. Dessa forma, pode-se dizer que o modelo não é robusto, visto que é sensível a medidas influentes.

3.5.3 Multicolinearidade

Para a verificação do modelo 4, será apresentado o quadro de correlação entre as variáveis. Além disso, é calculado o Fator de Inflação de Variância (VIF).

Quadro 15: Matriz de Correlação - Modelo 4

Correlação	ln (Mpg)	*Horsepower*	*Weight*	*Age*	*Japan*	*Europe*
ln $(Mpgl)$	1	-0,825	-0,877	-0,626	0,452	0,263
Horsepower	-0,825	1	0,859	0,456	-0,329	-0,253
Weight	-0,877	0,859	1	0,37	-0,443	-0,289
Age	-0,626	0,456	0,37	1	-0,196	-0,013
Japan	0,452	-0,329	-0,443	-0,196	1	-0,201
Europe	0,263	-0,253	-0,289	-0,013	-0,201	1

Quadro 16: Fator de Inflação de Variância - Modelo 4

Age	*Weight*	*Japan*	*Europe*	*Horsepower*
1,2832	4,6493	1,4821	1,2927	4,2313

Conforme o Quadro 15, é perceptível que existe uma correlação positiva forte entre *Weight* e *Horsepower*. Desse modo, há indícios de multicolinearidade entre algumas variáveis.

Mediante o Quadro 16, é notório que há fatores de inflação de variância consideravelmente maiores que 1 (*Weight* e *Horsepower*). Nesse caso, a média é, aproximadamente, 2,587, o que pode indicar multicolinearidade. Logo, será considerado mais um modelo.

3.5.4 Interpretação do Modelo

$$\text{Modelo 4: ln } \widehat{(Mpg)}_i = 4,0862 - 0,0311 Age_i - 0,0002 Weight_i + 0,094 Japan_i + 0,0797 Europe_i - 0,0011 Horsepower_i \quad (42)$$

- Quando todas as variáveis assumem valor zero, é esperado que o log do consumo seja 4,0862. Vale ressaltar que, neste caso, não faz sentido interpretar o intercepto isoladamente, pois *Weight* e *Horsepower* devem assumir valores positivos.
- É esperado que o log do consumo diminua em 0,0311 para cada aumento em uma unidade em *Age*, quando as outras variáveis explicativas estão fixas.
- É esperado que o log do consumo diminua em 0,0002, para cada aumento em uma unidade em *Weight*, quando as outras variáveis explicativas estão fixas.
- É esperado que o log do consumo aumente em 0,094 quando o veículo é do Japão em relação aos veículos dos Estados Unidos, mantidas as outras variáveis explicativas fixas.
- É esperado que o log do consumo aumente em 0,0797 quando o veículo é da Europa em relação aos veículos dos Estados Unidos, mantidas as outras variáveis explicativas fixas.
- É esperado que o log do consumo diminua em 0,0011, para cada aumento em uma unidade em *Horsepower*, quando as outras variáveis explicativas estão fixas.

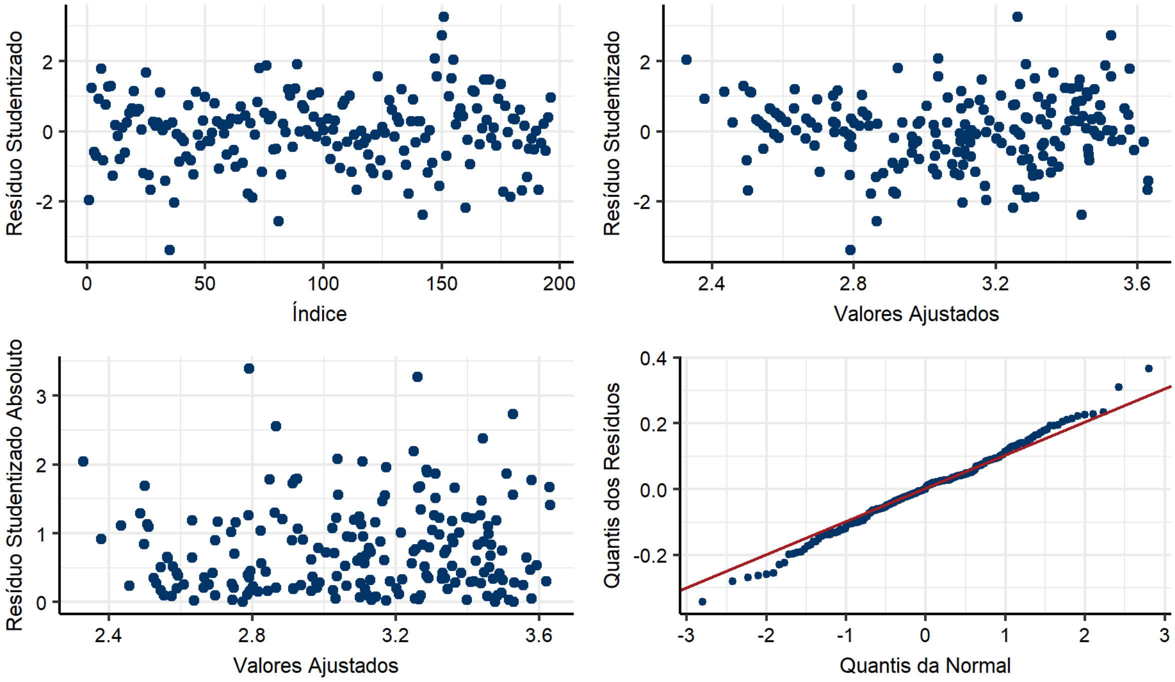
3.6 Modelo 3

Considerando que o Modelo 3 foi selecionado como modelo candidato, foi feita a análise de diagnóstico para verificar-se a adequação do modelo.

3.6.1 Verificação de pressupostos

Para a verificação dos pressupostos foram utilizados métodos gráficos e testes de hipótese.

Figura 13: Painel de verificação dos pressupostos



Levando em consideração a Figura 13, pode-se verificar os pressupostos do modelo 3. O gráfico dos resíduos *studentizados* pelo índice indica que há independência das observações, visto que apresenta aparente aleatoriedade no comportamento.

Os gráficos referentes aos valores ajustados pelos resíduos *studentizados* fornecem informação a respeito da homocedasticidade e linearidade do modelo. Nesse caso, nota-se que os gráficos estão de acordo com os dois pressupostos, visto que não parece haver nenhum tipo de padrão não aleatório no comportamento dos resíduos *studentizados*.

O gráfico de probabilidade normal verifica que há normalidade no comportamento dos resíduos *studentizados*, visto que não há um desvio significativo da linha representando os quantis da distribuição normal. Ou seja, é um indicio de que os resíduos do modelo 4 seguem uma distribuição normal.

Quadro 17: Testes de Hipótese dos pressupostos

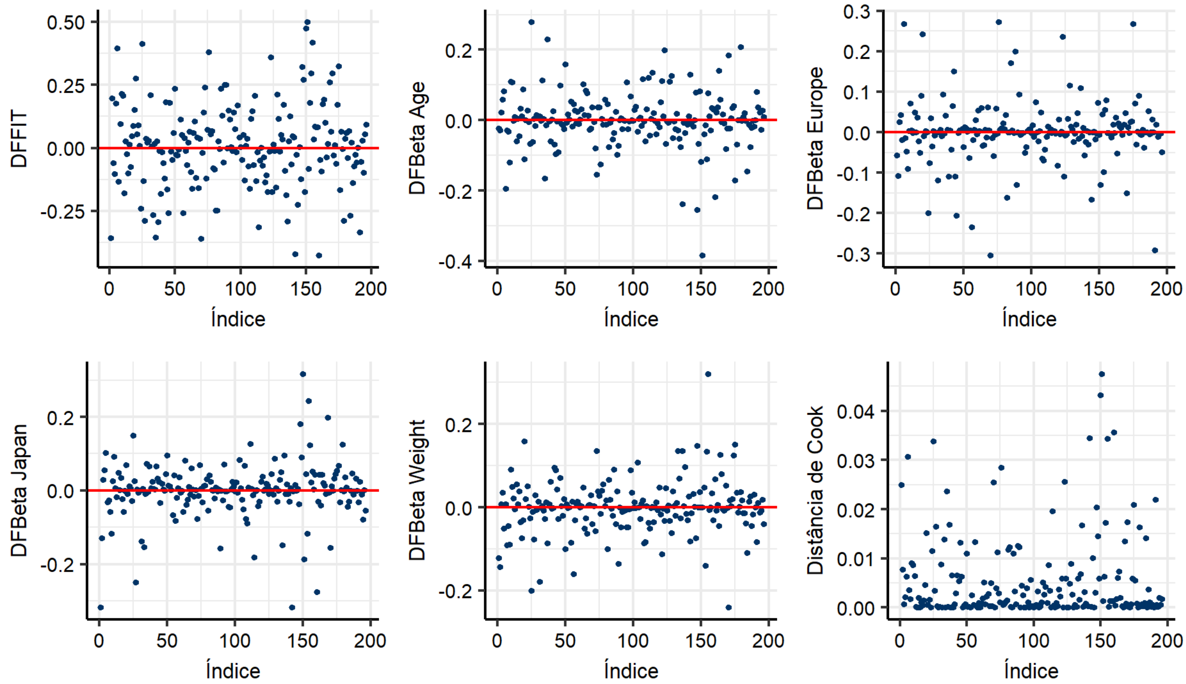
Teste	**Estatística do Teste**	**P-valor**
Breusch-Pagan	1,4447	0,8364
Shapiro-Wilk	0,9953	0,7958

Considerando as informações do Quadro 17 e também o nível de significância de $\alpha = 0,05$, pode-se concluir que o modelo 4 atende ao pressuposto de homocedasticidade, visto que o teste de *Breusch-Pagan* possui p-valor de 0,8364, maior que α. O pressuposto de normalidade também é atendido, uma vez que o teste de Shapiro-Wilk tem p-valor de 0,7958, maior que α.

3.6.2 Medidas influentes

Deseja-se verificar se o modelo 3 possui alguma estimativa que é influenciada por valores extremos nas observações. Para isso, serão utilizadas as medidas $DFBeta$, $DFFIT$ e a Distância de Cook.

Figura 14: Medidas influentes do modelo 3



Ao observar os gráficos, as observações que podem ser consideradas influentes são 150 e 151. Foi realizada uma nova regressão retirando essas observações, considerando todas as combinações entre elas. Depois dessa análise, não foi verificado nenhum valor discrepante a ponto de ser considerado influente, o que indica que o modelo 3 é um modelo de regressão robusto.

3.6.3 Multicolinearidade

Quadro 18: Fator de Inflação de Variância - Modelo 3

Age	*Weight*	*Japan*	*Europe*
1,1723	1,7184	1,4599	1,2927

Mediante o Quadro 15, nota-se que, como *Horsepower* já não está mais no modelo, existem apenas correlações fracas entre as variáveis explicativas.

Consoante o Quadro 18, não existem fatores de inflação drasticamente maiores que 1. Além disso, nesse caso, a média é de aproximadamente 1,4108. Portanto, é seguro afirmar que multicolinearidade não está afetando significativamente este modelo.

3.6.4 Interpretação do Modelo

$$\text{Modelo 3: } \ln\widehat{(Mpg)}_i = 4,1053 - 0,0329Age_i - 0,0003Weight_i + 0,0866Japan_i + 0,0792Europe_i$$

- Quando todas as variáveis assumem valor zero, é esperado que o log do consumo seja 4,1053. Vale ressaltar que, neste caso, não faz sentido interpretar o intercepto isoladamente, pois *Weight* deve assumir valor positivo.
- É esperado que o log do consumo diminua em 0,0329, para cada aumento em uma unidade em *Age*, quando as outras variáveis explicativas estão fixas.
- É esperado que o log do consumo diminua em 0,0003, para cada aumento em uma unidade em *Weight*, quando as outras variáveis explicativas estão fixas.
- É esperado que o log do consumo aumente em 0,0866 quando o veículo é do Japão, em relação aos veículos dos Estados Unidos, mantidas as outras variáveis explicativas fixas.
- É esperado que o log do consumo aumente em 0,0792 quando o veículo é da Europa em relação aos veículos dos Estados Unidos, mantidas as outras variáveis explicativas fixas.

4 Resultados - Amostra de Validação

Para a verificação da adequabilidade dos modelos encontrados, será realizada uma análise das mudanças nas estimativas dos parâmetros, do erro de predição quadrático médio (MSPR) e capacidade de predição do modelo a partir de um intervalo de predição. Nesta seção, utilizou-se um nível de significância de 5%.

4.1 Modelo 4

4.1.1 Coeficientes

São apresentados os intervalos de confiança e a estimativa pontual dos coeficientes do modelo 4 encontrado na etapa de desenvolvimento. Além disso, é exibida a estimativa pontual da etapa de validação, com o intuito de realizar a comparação.

Quadro 19: Comparação de Coeficientes - Modelo 4

Variável	Inferior	Superior	Pontual - Desenvolvimento	Pontual - Validação
Intercept	4,0007	4,1717	4,0862	4,1856
Age	-0,0362	-0,026	-0,0311	-0,0295
Weight	-0,0003	-0,0002	-0,0002	-0,0003
Japan	0,0423	0,1458	0,094	0,034
Europe	0,0264	0,1329	0,0797	0,0555
Horsepower	-0,0019	-0,0002	-0,0011	-0,0007

Conforme o Quadro 19 apresentado, nota-se que, para todas as variáveis, com exceção de *Japan* e do intercepto, as estimativas pontuais da etapa de validação encontram-se dentro do intervalo de confiança de 95%. Ademais, as estimativas pontuais dessas duas etapas não se distanciam drasticamente, com exceção das variáveis *Japan* e *Europe*.

4.1.2 Erro de Predição Quadrático Médio

Foi encontrado o valor do Erro de Predição Quadrático Médio de 0,0148, enquanto o quadrado médio do resíduo do modelo 4 foi de 0,0139, ou seja, como os valores estão próximos, há indícios de que o quadrado médio do modelo 4 não é seriamente viesado e o modelo 4 tem um bom poder preditivo.

4.1.3 Proporção de Aceitação

Com um intervalo de predição de 95% foi encontrada uma taxa de aceitação de 0,9286, ou seja, para a amostra de validação, apenas 7,14% dos valores observados estão fora do intervalo de predição, ao nível de significância de 5%.

4.2 Modelo 3

São apresentados os intervalos de confiança e a estimativa pontual dos coeficientes do modelo 3 encontrado na etapa de desenvolvimento. Além disso, é exibida a estimativa pontual da etapa de validação, com o intuito de realizar a comparação.

4.2.1 Coeficientes

Quadro 20: Comparação de Coeficientes - Modelo 3

Variável	Inferior	Superior	Pontual - Desenvolvimento	Pontual - Validação
Intercept	4,0205	4,1902	4,1053	4,2066
Age	-0,0378	-0,028	-0,0329	-0,0309
Weight	-0,0003	-0,0002	-0,0003	-0,0003
Japan	0,0347	0,1386	0,0866	0,0273
Europe	0,0253	0,1331	0,0792	0,0589

Conforme o Quadro 20, nota-se que, para todas as variáveis, as estimativas pontuais da etapa de validação encontram-se dentro do intervalo de confiança de 95%, com exceção de *Japan* e do intercepto. Ademais, as estimativas pontuais dessas duas etapas não se distanciam drasticamente, com exceção das variáveis *Japan* e *Europe*.

4.2.2 Erro de Predição Quadrático Médio

Foi encontrado o valor do Erro de Predição Quadrático Médio de 0,0151, enquanto o quadrado médio do resíduo do modelo 3 foi de 0,0143, ou seja, como os valores estão próximos, há indícios de que o quadrado médio do modelo 3 não é seriamente viesado e o modelo 3 tem um bom poder preditivo.

4.2.3 Proporção de Aceitação

Com um intervalo de predição de 95% foi encontrada uma taxa de aceitação de 0,9286, ou seja, para a amostra de validação, apenas 7,14% dos valores observados estão fora do intervalo de predição, ao nível de significância de 5%.

5 Conclusão

Mediante os resultados apresentados, obteve-se o modelo 3 e o modelo 4 como os mais adequados.

Em relação ao modelo 4, verificou-se que todos os pressupostos são válidos e que possui R^2_{adj} elevado, isto é, grande parte da variável resposta está sendo explicada pelo modelo. Contudo, foram observados indícios de multicolinearidade, o que pode comprometer a adequabilidade do modelo. Além disso, ele se apresentou sensível a observações discrepantes, ou seja, não aparenta ser robusto.

Em relação ao modelo 3, todos os pressupostos também são válidos e possui R^2_{ajd} elevado. Por sua vez, não foram encontradas evidências expressivas de multicolinearidade, o que mostra que ele não é influenciado pelas correlações entre as variáveis explicativas. Além disso, ele se apresentou robusto, pois não foi afetado significativamente por valores discrepantes.

Sendo assim, o modelo 3 possui vantagens em relação ao modelo 4.

Referências

Conover, W. J., & Conover, W. J. (1980). Practical nonparametric statistics.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1988). *Applied regression analysis and other multivariable methods* (Vol. 601). Duxbury Press Belmont, CA.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, *15*(4), 143–156.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., et al. (2005). *Applied linear statistical models* (Vol. 5). McGraw-Hill Irwin Boston.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (Vol. 4). Irwin Chicago.

Quinlan, J. R. (1993). Combining instance-based and model-based learning. In *Icml* (p. 236).

Ramos, E., & Donoho, D. (1983). Asa data exposition dataset. *CMU Dataset Archive*.

A Código R

```
require(ggplot2)
require(leaps)
require(readr)
require(MASS)
require(cowplot)
require(car)
options(scipen = 999)
auto_mpg <- read_table("~/UNB/5º Semestre/Análise de Regressão Linear/Trabalho/auto-mpg.data", col_names = FALSE)
auto_mpg$X9 <- sub("([0-9])\\t.*", "\\1", auto_mpg$X8)
auto_mpg$X8 <- sub('"',"",sub('([0-9])\\t\\"(.*)', "\\2", auto_mpg$X8))
str(auto_mpg)
auto_mpg <- auto_mpg[auto_mpg$X4 != "?",]
auto_mpg$X4 <- as.numeric(auto_mpg$X4); auto_mpg$X9 <- as.numeric(auto_mpg$X9)
str(auto_mpg)
auto_mpg <- data.frame(auto_mpg)
names(auto_mpg) <- c("mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration", "model year", "car name", "origin")
auto_mpg$`model year` <- 83 - auto_mpg$`model year`; names(auto_mpg)[7] <- "age"
auto_mpg$Japan <- as.numeric(auto_mpg$origin == 3); auto_mpg$Europe <- as.numeric(auto_mpg$origin == 2)
auto_mpg <- auto_mpg[,-c(8,9)]
set.seed(69)
id <- sample(1:nrow(auto_mpg), size = nrow(auto_mpg)/2)
desenv <- auto_mpg[id,]
valid <- auto_mpg[-id,]
desenv$origin <- NA
for(i in 1:196) {
  if(desenv$Japan[i] == 1) {desenv$origin[i] <- "Japão"} else if (desenv$Europe[i] == 1) {desenv$origin[i] <- "Europa"} else {desenv$origin[i] <- "EUA"}
}
```

```
desenv$origin <- factor(desenv$origin)

# Descritiva Univariada
quadro <- data.frame(matrix(NA, 7, 8))
names(quadro) <- c("Variável", "Mínimo", "1º Quartil", "Mediana", "Média", "3º
    Quartil", "Máximo", "Desvio padrão")
for(i in 1:7){
  nome <- names(desenv)[i]
  valores <- round(summary(desenv[[i]]),2)
  dp <- round(sd(desenv[[i]]),2)
  quadro[i,1] <- nome
  for(j in 2:8){
    quadro[i,j] <- c(valores,dp)[j-1]
  }
}
a <- c()
for(i in 1:7){
  a[i] <- paste(sub("\\.",",",paste(quadro[i,])), collapse = " & ")
}
paste(a, collapse = " \\ ")
names(desenv)
mpg <- ggplot(desenv, aes(x=factor(""), y=mpg)) +
  geom_boxplot(fill=c("#003366"), width = 0.5) +
  guides(fill=FALSE) +
  stat_summary(fun.y="mean", geom="point", shape=23, size=3, fill="white")+
  labs(x="", y="Consumo (milhas/galão)")+
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=12),
        axis.title.x = element_text(colour="black", size=12),
        axis.text = element_text(colour = "black", size=9.5),
        panel.border = element_blank(),
        axis.line.y = element_line(colour = "black"))
cyl <- ggplot(desenv, aes(x = cylinders)) + geom_bar(fill="#003366") +
  labs(x="Nº de cilindros", y="Frequência") +
```

```
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=12),
        axis.title.x = element_text(colour="black", size=12),
        axis.text = element_text(colour = "black", size=9.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black")) +
  scale_x_continuous(breaks = 3:8)
dis <- ggplot(desenv, aes(x=factor(""), y=displacement)) +
  geom_boxplot(fill=c("#003366"), width = 0.5) +
  guides(fill=FALSE) +
  stat_summary(fun.y="mean", geom="point", shape=23, size=3, fill="white")+
  labs(x="", y="Deslocamento (pol³)")+
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=12),
        axis.title.x = element_text(colour="black", size=12),
        axis.text = element_text(colour = "black", size=9.5),
        panel.border = element_blank(),
        axis.line.y = element_line(colour = "black"))
hor <- ggplot(desenv, aes(x=factor(""), y=horsepower)) +
  geom_boxplot(fill=c("#003366"), width = 0.5) +
  guides(fill=FALSE) +
  stat_summary(fun.y="mean", geom="point", shape=23, size=3, fill="white")+
  labs(x="", y="Potência (cv)")+
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=12),
        axis.title.x = element_text(colour="black", size=12),
        axis.text = element_text(colour = "black", size=9.5),
        panel.border = element_blank(),
        axis.line.y = element_line(colour = "black"))
wei <- ggplot(desenv, aes(x=factor(""), y=weight)) +
  geom_boxplot(fill=c("#003366"), width = 0.5) +
  guides(fill=FALSE) +
  stat_summary(fun.y="mean", geom="point", shape=23, size=3, fill="white")+
  labs(x="", y="Peso (libras)")+
```

```
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=12),
        axis.title.x = element_text(colour="black", size=12),
        axis.text = element_text(colour = "black", size=9.5),
        panel.border = element_blank(),
        axis.line.y = element_line(colour = "black"))
acc <- ggplot(desenv, aes(x=factor(""), y=acceleration)) +
  geom_boxplot(fill=c("#003366"), width = 0.5) +
  guides(fill=FALSE) +
  stat_summary(fun.y="mean", geom="point", shape=23, size=3, fill="white")+
  labs(x="", y="Aceleração (0-60 milhas)")+
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=12),
        axis.title.x = element_text(colour="black", size=12),
        axis.text = element_text(colour = "black", size=9.5),
        panel.border = element_blank(),
        axis.line.y = element_line(colour = "black"))
age <- ggplot(desenv, aes(x=factor(""), y=age)) +
  geom_boxplot(fill=c("#003366"), width = 0.5) +
  guides(fill=FALSE) +
  stat_summary(fun.y="mean", geom="point", shape=23, size=3, fill="white")+
  labs(x="", y="Idade em 1983")+
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=12),
        axis.title.x = element_text(colour="black", size=12),
        axis.text = element_text(colour = "black", size=9.5),
        panel.border = element_blank(),
        axis.line.y = element_line(colour = "black"))
ori <- ggplot(desenv, aes(x = origin)) + geom_bar(fill="#003366", width = 0.5)
    +
  labs(x="Origem", y="Frequência") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=12),
        axis.title.x = element_text(colour="black", size=12),
```

```
        axis.text = element_text(colour = "black", size=9.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
plot_grid(mpg, dis, hor, nrow = 1)
ggsave("univariada1.png", width = 158, height = 93, units = "mm")
plot_grid(wei,acc, age, nrow = 1)
ggsave("univariada2.png", width = 158, height = 93, units = "mm")
plot_grid(cyl, ori, nrow = 1)
ggsave("univariada3.png", width = 158, height = 93, units = "mm")

# Descritiva Bivariada
cyl <- ggplot(desenv, aes(x=cylinders, y=mpg)) + geom_point(colour="#003366",
    size=1.5) +
  labs(x="Nº de cilindros", y="Consumo (milhas/galão)") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
dis <- ggplot(desenv, aes(x=displacement, y=mpg)) + geom_point(colour="#003366"
    , size=1.5) +
  labs(x="Deslocamento (pol³)", y="Consumo (milhas/galão)") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
hor <- ggplot(desenv, aes(x=horsepower, y=mpg)) + geom_point(colour="#003366",
    size=1.5) +
  labs(x="Potência (cv)", y="Consumo (milhas/galão)") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
```

```
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
wei <- ggplot(desenv, aes(x=weight, y=mpg)) + geom_point(colour="#003366", size
    =1.5) +
  labs(x="Peso (libras)", y="Consumo (milhas/galão)") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black")) +
  scale_x_continuous(breaks = seq(2000, 5000, 1500))
acc <- ggplot(desenv, aes(x=acceleration, y=mpg)) + geom_point(colour="#003366"
    , size=1.5) +
  labs(x="Aceleração (0-60 milhas)", y="Consumo (milhas/galão)") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
age <- ggplot(desenv, aes(x=age, y=mpg)) + geom_point(colour="#003366", size
    =1.5) +
  labs(x="Idade em 1983", y="Consumo (milhas/galão)") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
ori <- ggplot(desenv, aes(x=origin, y=mpg)) +
  geom_boxplot(fill=c("#003366"), width = 0.5) +
```

```
  stat_summary(fun.y="mean", geom="point", shape=23, size=3, fill="white")+
  labs(x="Origem", y="Consumo (milhas/galão)") +
  stat_summary(geom = "crossbar", width=0.5, fatten=0, color="red",
               fun.data = function(x){ return(c(y=median(x), ymin=median(x), ymax
                   =median(x))) }) +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line.y = element_line(colour = "black"))
plot_grid(cyl, dis, hor, wei, nrow = 1)
ggsave("bivariada1.png", width = 158, height = 93, units = "mm")
plot_grid(acc, age, ori, nrow = 1)
ggsave("bivariada2.png", width = 158, height = 93, units = "mm")
r2 <- numeric(7)
for(i in 1:7){
  r2[i] <- 1-mean(tapply(desenv[[i]], desenv$origin, var))/var(desenv[[i]])
}
names(r2) <- names(desenv)[1:7]
paste(sub("\\.",",",paste(round(r2,3))), collapse = " & ")
desenv <- desenv[,-c(10)]
cor <- data.frame(cor(desenv)[1:7,1:7])
for(i in 1:7){
  print(gsub("\\.",",",paste(round(cor[i,],3), collapse = " & ")))
}
cor(desenv)
plot(desenv)

# Modelo
mod <- lm(mpg ~ ., data = desenv)
summary(mod)
# Resíduo
res <- mod$residuals
```

```r
# Resíduo studentizado
res_stud <- rstudent(mod)
# Independência
plot(res_stud)
df <- data.frame(indice = 1:196, res, res_stud, fitt = mod$fitted.values)
ind <- ggplot(df, aes(x=indice, y=res_stud)) + geom_point(colour="#003366",
    size=1.5) +
  labs(x="Índice", y="Resíduo Studentizado") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
# Homocedasticidade e Linearidade
plot(mod$fitted.values, res_stud)
lin <- ggplot(df, aes(x=fitt, y=res_stud)) + geom_point(colour="#003366", size
    =1.5) +
  labs(x="Valores Ajustados", y="Resíduo Studentizado") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
plot(mod$fitted.values, abs(res_stud))
linabs <- ggplot(df, aes(x=fitt, y=abs(res_stud))) + geom_point(colour="#003366
    ", size=1.5) +
  labs(x="Valores Ajustados", y="Resíduo Studentizado Absoluto") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
```

```r
       axis.line = element_line(colour = "black"))
# Normalidade
boxplot(res_stud)
qqnorm(res_stud)
qqline(res_stud)

y <- quantile(res, c(0.25, 0.75))
x <- qnorm(c(0.25,0.75))
slope <- diff(y)/diff(x)
int <- y[1L] - slope * x[1L]
d <- data.frame(resids = res)
qq <- ggplot(d, aes(sample = resids)) + stat_qq(colour = "#003366", size = 1) +
  geom_abline(slope = slope, intercept = int, size = .5, colour = "#A11D21")+
  xlab("Quantis da Normal")+ylab("Quantis dos Resíduos") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
       axis.title.x = element_text(colour="black", size=8),
       axis.text = element_text(colour = "black", size=7.5),
       panel.border = element_blank(),
       axis.line = element_line(colour = "black"))
plot_grid(ind, lin, linabs, qq, nrow = 2)
ggsave("diagnostico_completo.png", width = 158, height = 93, units = "mm")
# Testes
shapiro.test(res)
modbp <- lm(((res)^2) ~ . - mpg, data = desenv)
SQReg <- sum(anova(modbp)[1:8,2])
SQRes <- anova(mod)[9,2]
testchi <- (SQReg/9)/((SQRes/length(desenv$mpg))^2)
(pvalor <- 1-pchisq(testchi, 8))

# Transformação: log(mpg)
desenv$mpgl <- log(desenv$mpg)
desenv <- desenv[,-1]
# Modelo
```

```r
mod <- lm(mpgl ~ ., data = desenv)
summary(mod)
# Resíduo
res <- mod$residuals
# Resíduo studentizado
res_stud <- rstudent(mod)
# Independência
plot(res_stud)
df <- data.frame(indice = 1:196, res, res_stud, fitt = mod$fitted.values)
ind <- ggplot(df, aes(x=indice, y=res_stud)) + geom_point(colour="#003366",
    size=1.5) +
  labs(x="Índice", y="Resíduo Studentizado") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
# Homocedasticidade e Linearidade
plot(mod$fitted.values, res_stud)
lin <- ggplot(df, aes(x=fitt, y=res_stud)) + geom_point(colour="#003366", size
    =1.5) +
  labs(x="Valores Ajustados", y="Resíduo Studentizado") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
plot(mod$fitted.values, abs(res_stud))
linabs <- ggplot(df, aes(x=fitt, y=abs(res_stud))) + geom_point(colour="#003366
    ", size=1.5) +
  labs(x="Valores Ajustados", y="Resíduo Studentizado Absoluto") +
  theme_bw() +
```

```
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
# Normalidade
boxplot(res_stud)
qqnorm(res_stud)
qqline(res_stud)

y <- quantile(res, c(0.25, 0.75))
x <- qnorm(c(0.25,0.75))
slope <- diff(y)/diff(x)
int <- y[1L] - slope * x[1L]
d <- data.frame(resids = res)
qq <- ggplot(d, aes(sample = resids)) + stat_qq(colour = "#003366", size = 1) +
  geom_abline(slope = slope, intercept = int, size = .5, colour = "#A11D21")+
  xlab("Quantis da Normal")+ylab("Quantis dos Resíduos") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black"))
plot_grid(ind, lin, linabs, qq, nrow = 2)
ggsave("diagnostico_transformado.png", width = 158, height = 93, units = "mm")
# Testes
shapiro.test(res)
modbp <- lm(((res)^2) ~ . - mpgl, data = desenv)
SQReg <- sum(anova(modbp)[1:8,2])
SQRes <- anova(mod)[9,2]
testchi <- (SQReg/9)/((SQRes/length(desenv$mpgl))^2)
(pvalor <- 1-pchisq(testchi, 8))
```

```
# Fazer seleção de variáveis
# Medidas: R2, R2adj, BIC, Cm
sele1 <- regsubsets(mpgl ~ ., data = desenv, nbest = 10)
names(summary(sele1))
n_var_exp <- as.numeric(rownames(summary(sele1)$which))
df <- data.frame(n_var_exp, r2 = summary(sele1)$rsq, r2adj = summary(sele1)$
    adjr2, bic = summary(sele1)$bic, cm = summary(sele1)$cp)
plot(n_var_exp, summary(sele1)$rsq, xlab = "Nº de variáveis explicativas", ylab
     = "R^2") # 2, 3, 4
r2 <- ggplot(df, aes(x=n_var_exp, y=r2)) + geom_point(colour="#003366", size
    =1.5, alpha = 0.5) +
  labs(x="Número de variáveis explicativas", y="R²") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black")) +
  scale_x_continuous(breaks = 1:8)
plot(n_var_exp, summary(sele1)$adjr2, xlab = "Nº de variáveis explicativas",
    ylab = "R^2adj") # 2, 3, 4, 5, 6
r2adj <- ggplot(df, aes(x=n_var_exp, y=r2adj)) + geom_point(colour="#003366",
    size=1.5, alpha = 0.5) +
  labs(x="Número de variáveis explicativas", y="R² Ajustado") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black")) +
  scale_x_continuous(breaks = 1:8)
plot(n_var_exp, summary(sele1)$bic, xlab = "Nº de variáveis explicativas", ylab
     = "BIC") # 2, 3, 4
bic <- ggplot(df, aes(x=n_var_exp, y=bic)) + geom_point(colour="#003366", size
```

```
    =1.5, alpha = 0.5) +
  labs(x="Número de variáveis explicativas", y="BIC") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black")) +
  scale_x_continuous(breaks = 1:8)
plot(n_var_exp, summary(sele1)$cp, xlab = "Nº de variáveis explicativas", ylab
    = "Cm") # 2, 3, 4
cm <- ggplot(df, aes(x=n_var_exp, y=cm)) + geom_point(colour="#003366", size
    =1.5, alpha = 0.5) +
  labs(x="Número de variáveis explicativas", y="C de Mallows") +
  theme_bw() +
  theme(axis.title.y=element_text(colour="black", size=8),
        axis.title.x = element_text(colour="black", size=8),
        axis.text = element_text(colour = "black", size=7.5),
        panel.border = element_blank(),
        axis.line = element_line(colour = "black")) +
  scale_x_continuous(breaks = 1:8)
plot_grid(r2, r2adj, bic, cm, nrow = 2)
ggsave("selecao.png", width = 158, height = 93, units = "mm")
mod1 <- lm(mpgl ~ age + weight, data = desenv) # 2 variáveis
mod2 <- lm(mpgl ~ age + weight + Japan, data = desenv) # 3 variáveis
mod3 <- lm(mpgl ~ age + weight + Japan + Europe, data = desenv) # 4 variáveis
mod4 <- lm(mpgl ~ age + weight + Japan + Europe + horsepower, data = desenv) #
    5 variáveis
mod5 <- lm(mpgl ~ age + weight + Japan + Europe + horsepower + displacement,
    data = desenv) # 6 variáveis
r <- round(cbind(summary(sele1)$which, summary(sele1)$rsq, summary(sele1)$adjr2
    , summary(sele1)$bic, summary(sele1)$cp),4)
r <- r[c(9,19,29,39,49,59),10:13]
colnames(r) <- names(summary(sele1))[c(2,4,5,6)]
```

```
b <- c()
for(i in 1:6){
  b[i] <- paste(i, paste(r[i,], collapse = " "))
}
gsub("\\.",",",b)
# Implementar Backward
summary(lm(mpgl ~ ., data = desenv)) # Retira-se acceleration
summary(lm(mpgl ~ . - acceleration, data = desenv)) # Retira-se cylinders
summary(lm(mpgl ~ . - acceleration - cylinders, data = desenv)) # Retira-se
    displacement
summary(lm(mpgl ~ . - acceleration - cylinders - displacement, data = desenv))
    # Modelo final igual ao mod4
# Métodos automáticos
full.model <- lm(mpgl ~ ., data = desenv)
back.model <- stepAIC(full.model, direction = "backward", trace = F)
summary(back.model)
mod6 <- lm(mpgl ~ age + weight + Japan + Europe + horsepower + displacement +
    cylinders, data = desenv)
forw.model <- stepAIC(lm(mpgl ~ 1, data=desenv), direction="forward", scope=(~
    cylinders + displacement + horsepower + weight +
                                                                    acceleration
                                                                       +
                                                                      age
                                                                      +
                                                                      Japan
                                                                       +
                                                                      Europe
                                                                      ),
                        trace=F)
summary(forw.model) # Modelo final igual ao mod4
step.model <- stepAIC(lm(mpgl ~ 1, data=desenv), direction="both", scope=list(
    lower=lm(mpgl ~ 1, data=desenv), upper = full.model), trace=F)
summary(step.model) # Modelo final igual ao mod4
# Definindo os modelos candidatos
```

```r
summary(mod1)
summary(mod2)
summary(mod3)
summary(mod4)
summary(mod5) # Descarta-se o mod5, pois displacement não é significativo
summary(mod6) # Descarta-se o mod6, pois displacement e cylinders não são
    significativos

# Análise completa de diagnóstico do mod4
desenv2 <- desenv[,c(3,4,6:9)]
# Os Pressupostos estão validados?
summary(mod4)
# Resíduo
res <- mod4$residuals
# Resíduo studentizado
res_stud <- rstudent(mod4)
# Independência
plot(res_stud)
# Homocedasticidade e Linearidade
plot(mod4$fitted.values, res_stud)
plot(mod4$fitted.values, abs(res_stud))
# Normalidade
boxplot(res_stud)
qqnorm(res_stud)
qqline(res_stud)
# Testes
shapiro.test(res)
modbp <- lm(((res)^2) ~ . - mpgl, data = desenv2)
SQReg <- sum(anova(modbp)[1:5,2])
SQRes <- anova(mod)[6,2]
testchi <- (SQReg/6)/((SQRes/length(desenv2$mpgl))^2)
(pvalor <- 1-pchisq(testchi, 5))

# Medidas Influentes
```

```
medidas <- as.data.frame(influence.measures(mod4)[[1]])
plot(1:196, medidas$hat)
identify(1:196, medidas$hat) # 136
plot(1:196, abs(medidas$dffit))
identify(1:196, abs(medidas$dffit)) # 35
plot(1:196, medidas$cook.d)
identify(1:196, medidas$cook.d) # 10, 35
plot(1:196, abs(medidas$dfb.age))
identify(1:196, abs(medidas$dfb.age)) # 25, 151
plot(1:196, abs(medidas$dfb.wght))
identify(1:196, abs(medidas$dfb.wght)) # 35, 68, 136
plot(1:196, abs(medidas$dfb.Japn))
identify(1:196, abs(medidas$dfb.Japn)) # 1, 27, 142, 150, 154, 160
plot(1:196, abs(medidas$dfb.Eurp))
identify(1:196, abs(medidas$dfb.Eurp)) # 6, 56, 70, 76, 191
plot(1:196, abs(medidas$dfb.hrsp))
identify(1:196, abs(medidas$dfb.hrsp)) # 10, 35, 46, 68, 136
# 10, 35, 68, 136
summary(lm(mpgl ~ ., data = desenv2))
summary(lm(mpgl ~ ., data = desenv2[-10,]))
summary(lm(mpgl ~ ., data = desenv2[-35,])) # Horsepower
summary(lm(mpgl ~ ., data = desenv2[-68,]))
summary(lm(mpgl ~ ., data = desenv2[-136,])) # Horsepower
summary(lm(mpgl ~ ., data = desenv2[-c(10,35),]))
summary(lm(mpgl ~ ., data = desenv2[-c(10,68),]))
summary(lm(mpgl ~ ., data = desenv2[-c(10,136),]))
summary(lm(mpgl ~ ., data = desenv2[-c(35,68),]))
summary(lm(mpgl ~ ., data = desenv2[-c(35,136),])) # Horsepower
summary(lm(mpgl ~ ., data = desenv2[-c(68,136),]))
summary(lm(mpgl ~ ., data = desenv2[-c(10,35,68),]))
summary(lm(mpgl ~ ., data = desenv2[-c(10,35,136),])) # Horsepower
summary(lm(mpgl ~ ., data = desenv2[-c(10,68,136),]))
summary(lm(mpgl ~ ., data = desenv2[-c(35,68,136),])) # Horsepower
summary(lm(mpgl ~ ., data = desenv2[-c(10,35,68,136),])) # Horsepower
```

```
# Multicolinearidade
(vi<-vif(mod4))
mean(vi)

# Análise completa de diagnóstico do mod3
desenv3 <- desenv[,c(4,6:9)]
# Os Pressupostos estão validados?
summary(mod3)
# Resíduo
res <- mod3$residuals
# Resíduo studentizado
res_stud <- rstudent(mod3)
# Independência
plot(res_stud)
# Homocedasticidade e Linearidade
plot(mod3$fitted.values, res_stud)
plot(mod3$fitted.values, abs(res_stud))
# Normalidade
boxplot(res_stud)
qqnorm(res_stud)
qqline(res_stud)
# Testes
shapiro.test(res)
modbp <- lm(((res)^2) ~ . - mpgl, data = desenv3)
SQReg <- sum(anova(modbp)[1:4,2])
SQRes <- anova(mod3)[5,2]
testchi <- (SQReg/5)/((SQRes/length(desenv3$mpgl))^2)
(pvalor <- 1-pchisq(testchi, 4))

medidas <- as.data.frame(influence.measures(mod3)[[1]])
plot(1:196, medidas$hat)
identify(1:196, medidas$hat) # 126
plot(1:196, abs(medidas$dffit))
```

```
identify(1:196, abs(medidas$dffit)) # 150, 151
plot(1:196, medidas$cook.d)
identify(1:196, medidas$cook.d) # 150, 151
plot(1:196, abs(medidas$dfb.age))
identify(1:196, abs(medidas$dfb.age)) # 151
plot(1:196, abs(medidas$dfb.wght))
identify(1:196, abs(medidas$dfb.wght)) # 155, 170
plot(1:196, abs(medidas$dfb.Japn))
identify(1:196, abs(medidas$dfb.Japn)) # 1, 142, 150
plot(1:196, abs(medidas$dfb.Eurp))
identify(1:196, abs(medidas$dfb.Eurp)) # 70, 191
# 150, 151
summary(lm(mpgl ~ ., data = desenv3))
summary(lm(mpgl ~ ., data = desenv3[-150,]))
summary(lm(mpgl ~ ., data = desenv3[-151,]))
summary(lm(mpgl ~ ., data = desenv3[-c(150,151),]))

# Multicolinearidade
(vi<-vif(mod3))
mean(vi)



# Validação do modelo
# Modelo: mod4
valid$mpgl <- log(valid$mpg); valid$mpg <- NULL
mod4_valid <- lm(mpgl ~ age + weight + Japan + Europe + horsepower, data =
    valid)
summary(mod4_valid) # Japão e Horsepower deixaram de ser significativos
summary(mod4)
coef(mod4_valid)
coef(mod4) # Intercepto, Age, Weight e Horsepower são próximos. Japão e Europa
    são distantes, mas pouco.

confint(mod4, level = .99) #Todos dentro a 99%
```

```
# Intervalo de confiança de 95%
confianca_mod4 <- predict(mod4, valid[,-9], interval = "confidence", level =
    .95)
confidence_rate_mod4 <- c()
for(i in seq_len(nrow(confianca_mod4))){
  confidence_rate_mod4[i] <- valid[i,"mpgl"] >= confianca_mod4[i,"lwr"] & valid
      [i,"mpgl"] <= confianca_mod4[i,"upr"]
}
mean(confidence_rate_mod4)

predicao_mod4 <- predict(mod4, valid[,-9], interval = "prediction", level =
    .95)
predict_rate_mod4 <- c()
for(i in seq_len(nrow(confianca_mod4))){
  predict_rate_mod4[i] <- valid[i,"mpgl"] >= predicao_mod4[i,"lwr"] & valid[i,"
      mpgl"] <= predicao_mod4[i,"upr"]
}
mean(predict_rate_mod4)

mpgl_hat <- predict(mod4, valid[,-9])
MSPR <- mean((valid[,"mpgl"] - mpgl_hat)^2)
anova(mod4)["Residuals","Mean Sq"]
MSPR # Estão próximos

# Modelo: mod3
mod3_valid <- lm(mpgl ~ age + weight + Japan + Europe, data = valid)
summary(mod3_valid) # Japão deixou de ser significativo
summary(mod3)
coef(mod3_valid)
coef(mod3) # Intercepto, Age, Weight e Europa são próximos. Japão está distante
    .

confint(mod3, level = .99)
```

```r
# Intervalo de confiança de 95%
confianca_mod3 <- predict(mod3, valid[,-9], interval = "confidence", level =
    .95)
confidence_rate_mod3 <- c()
for(i in seq_len(nrow(confianca_mod3))){
  confidence_rate_mod3[i] <- valid[i,"mpgl"] >= confianca_mod3[i,"lwr"] & valid
      [i,"mpgl"] <= confianca_mod3[i,"upr"]
}
mean(confidence_rate_mod3)

predicao_mod3 <- predict(mod3, valid[,-9], interval = "prediction", level =
    .95)
predict_rate_mod3 <- c()
for(i in seq_len(nrow(confianca_mod3))){
  predict_rate_mod3[i] <- valid[i,"mpgl"] >= predicao_mod3[i,"lwr"] & valid[i,"
      mpgl"] <= predicao_mod3[i,"upr"]
}
mean(predict_rate_mod3)

mpgl_hat <- predict(mod3, valid[,-9])
MSPR <- mean((valid[,"mpgl"] - mpgl_hat)^2)
anova(mod3)["Residuals","Mean Sq"]
MSPR # Estão próximos
```