Atividade 9

Transformações e seleção de variáveis

Paulo Ricardo Seganfredo Campana

27 de outubro de 2023

Considerando novamente os dados da **Tabela B.3** do livro do *Montgomery*, sobre o consumo de combustível de diferentes automóveis, responda as questões abaixo, fixando o nível de significância em 5%.

Questão 1

a) Estime um modelo de regressão linear que relaciona o consumo de combustível em milhas/galão, y, com o volume de deslocamento do motor (cilindrada), x_1.

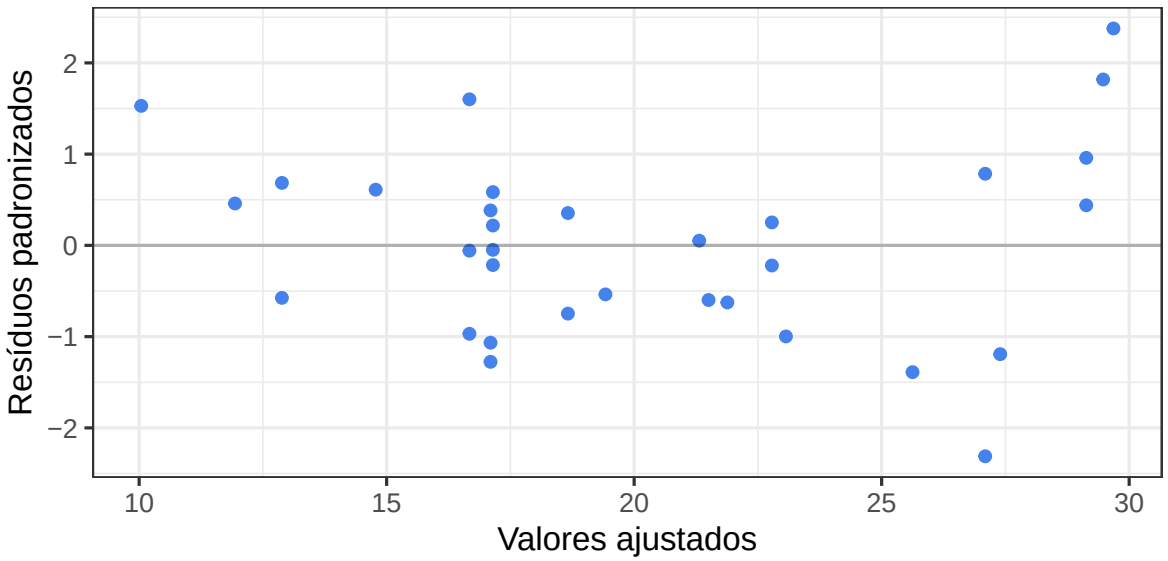
```
data <- MPV::table.b3
fit1 <- lm(y ~ x1, data)
```

$$\hat{y} = 33.723 - 0.047x_1$$

b) Verifique se o modelo ajustado é homoscedástico, se essa suposição for violada, proceda a correção das estimativas de Var(β) por meio dos HC's (utilize a função *vcovHC*, biblioteca *sandwich*). Comente os resultados;

```
lmtest::gqtest(fit1)$p.value
## [1] 0.723489
lmtest::bptest(fit1, studentize = FALSE)$p.value
##         BP
## 0.02672279
lmtest::bptest(fit1, studentize = TRUE)$p.value
##         BP
## 0.02059295
```

Figura 1: Gráfico para verificação da suposição de homoscedasticidade: valores ajudastos versus resíduos padronizados



É difícil notar algum padrão de mudança de variância dos resíduos dependente dos valores ajustados, possivelmente devido a baixa quantidade de observações, o teste de Goldfeld-Quandt não rejeita a suposição de homoscedasticidade porém os testes de Breusch-Pagan e Koenker rejeitam essa suposição.

c) Aplique a transformação de Box-Cox e verifique se a mesma consegue estabilizar a variância no modelo final. Comente os resultados;

```
bct <- car::powerTransform(fit1, family = "bcPower")
data$ybc <- car::bcPower(data$y, bct$lambda)
fit2 <- lm(ybc ~ x1, data)

lmtest::gqtest(fit2)$p.value
## [1] 0.28088
lmtest::bptest(fit2, studentize = FALSE)$p.value
##        BP
## 0.6192929
lmtest::bptest(fit2, studentize = TRUE)$p.value
##        BP
## 0.5720964
```

Sim, a transformação de Box-Cox com $\lambda = -0.218$ ajudou a estabilizar a variância dos resíduos, com isso todos os testes suportam a hipótese da homoscedasticidade do modelo.

Questão 2

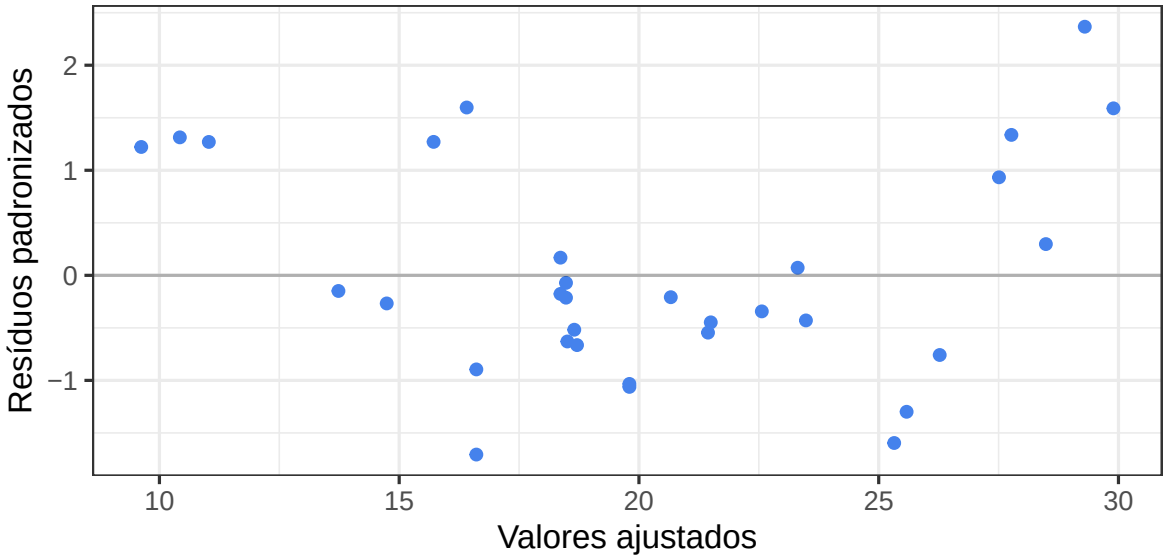
a) Ajuste um modelo de regressão linear que relaciona a variável consumo de combustível em milhas/galão, y, com o peso do carro, x_{10}.

```
fit3 <- lm(y ~ x10, data)
```

$$\hat{y} = 40.852 - 0.00575x_{10}$$

b) Construa um gráfico dos resíduos padronizados versus a resposta prevista. É possível identificar algum problema de adequação no modelo ao analisar este gráfico?

Figura 2: Gráfico para verificação da suposição de homoscedasticidade: valores ajudastos versus resíduos padronizados

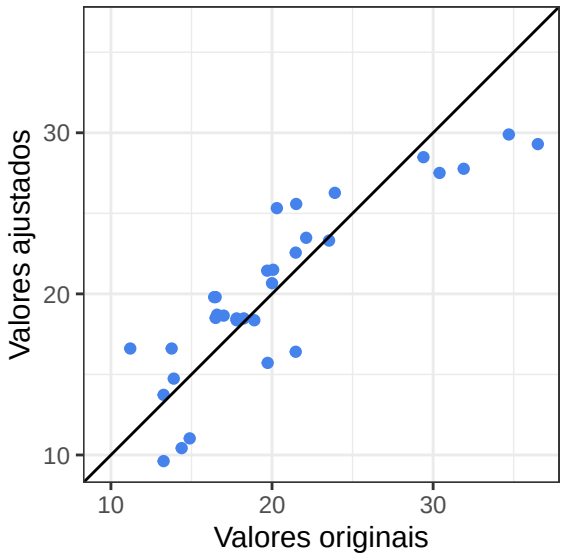


Sim, abaixo da reta $y = 0$, os resíduos padronizados são geralmente sobre valores de $\hat{y}$ medianos enquanto que acima da reta são valores mais extremos em ambas as direções, possivelmente indicando a não linearidade do modelo.

c) Faça um gráfico dos valores de y versus a resposta prevista. Parece que um modelo linear é adequado?

Um modelo linear possivelmente sim, porém esse modelo utilizando apenas x_{10} não, o mesmo falha em ajustar para valores de y altos, subestimando-os.

Figura 3: Gráfico para verificação da suposição de linearidade: Valores ajustados versus valores originais



d) Verifique a hipótese de linearidade. Quais são as conclusões sobre a adequação do modelo com relação a esse pressuposto?

```
lmtest::resettest(fit3)$p.value
## [1] 4.369778e-05
lmtest::raintest(fit3)$p.value
## [1] 0.9463242
```

O teste RESET rejeita a suposição de linearidade do modelo, enquanto que o teste *rainbow* não rejeita a mesma devido ao modelo linear ser suficientemente razoável. O não cumprimento dessa suposição nos leva a considerar transformações nas variáveis regressoras ou resposta, ou até mesmo um modelo não linear.

e) Verifique a hipótese de normalidade Quais são as conclusões sobre a adequação do modelo com relação a esse pressuposto?

```
res <- rstandard(fit3)
nortest::ad.test(res)$p.value
## [1] 0.02708782
nortest::cvm.test(res)$p.value
## [1] 0.01732475
nortest::lillie.test(res)$p.value
```

```
## [1] 0.04106568
nortest::pearson.test(res)$p.value
## [1] 0.01560942
nortest::sf.test(res)$p.value
## [1] 0.1097261
shapiro.test(res)$p.value
## [1] 0.0896049
moments::jarque.test(res)$p.value
## [1] 0.4237872
```

A maioria dos testes para normalidade dos resíduos rejeita esta suposição, com isso os intervalos de confiança e testes de hipótese para os coeficientes do modelo e predição não são confiáveis, pois partem do pressuposto de normalidade para o cálculo de estatísticas.

f) Utilizando o método de Box-Tidwell, identifique uma transformação apropriada para este caso. Ajuste o modelo com a variável transformada e verifique a adequação deste modelo. Comente os resultados.

```
BT <- car::boxTidwell(data$y ~ data$x10)
alphaBT <- BT$result[1]
alphaBT
## [1] -1.834878

data$x10BT <- data$x10 ^ alphaBT
fit4 <- lm(y ~ x10BT, data)
```

Com o método de Box-Tidwell obtemos $\alpha = -1.835$, essa transformação nos leva ao modelo

$$\hat{y} = 10.09 + 27435160 x_{10}^{-1.835}$$

```
# Normalidade
res <- rstandard(fit4)
nortest::lillie.test(res)$p.value
## [1] 0.6385049
shapiro.test(res)$p.value
## [1] 0.9819442
moments::jarque.test(res)$p.value
## [1] 0.9657394

# Linearidade
lmtest::resettest(fit4)$p.value
```

```
## [1] 0.9851509
lmtest::raintest(fit4)$p.value
## [1] 0.6765056

# Homoscedasticidade
lmtest::bgtest(fit4)$p.value
## [1] 0.2928148
lmtest::bptest(fit4, studentize = FALSE)$p.value
##        BP
## 0.9821601
lmtest::bptest(fit4, studentize = TRUE)$p.value
##       BP
## 0.982374

# Autocorrelação
lmtest::dwtest(fit4)$p.value
## [1] 0.8210621
```

O novo modelo com a transformação de Box-Tidwell é aceito por todos os testes de normalidade, linearidade, homoscedasticidade e autocorrelação, não é necessário verificar a suposição da multicolinearidade pois se trata de um modelo de regressão linear simples.

Questão 3

a) Sem considerar a variável x_3, use os métodos de seleção *forward* e de eliminação *backward* para especificar um modelo de regressão para explicar o consumo de gasolina por milhas por um subconjunto de variáveis regressoras. Ambos os procedimentos levaram ao mesmo modelo final? Expresse o modelo final.

```
semx3 <- MPV::table.b3[-4]
menor <- lm(y ~ 1, semx3)
maior <- lm(y ~ ., semx3)

foward <- step(menor, scope = list(lower = menor, upper = maior), direction = "forward")
backward <- step(maior, direction = "backward")
```

Não, pelo método forward obtemos o modelo $\hat{y} = 33.885 - 0.053x_1 + 0.959x_6$ enquanto que por eliminação backward o modelo é $\hat{y} = 5.011 + 2.625x_5 + 0.212x_8 - 0.0093x_{10}$.

b) Ajuste uma regressão linear múltipla que relaciona o consumo de combustível, y, com o volume de deslocamento do motor (cilindrada), x_1, e o número de carburadores, x_6.

Faça uma análise de variância e teste a significância global da regressão e escreva o que é possível concluir com este resultado? Faça o teste individual sobre os coeficientes da regressão, quais variáveis regressoras são significativas para o modelo? Apresente os resultados!

```
fit5 <- lm(y ~ x1 + x6, data)
```

Tabela 1: Análise de variância para o modelo utilizando x_1 e x_6

Variável	gl	SS	MS	Estatística	p.value
x_1	1	955.72	955.72	105.290	3.66×10^{-11}
x_6	1	18.59	18.59	2.048	0.163
Regressão	2	974.31	487.16	53.670	1.79×10^{-10}
Resíduos	29	263.235	9.077		

Tabela 2: Teste t para os coeficientes da regressão

Coeficiente	Estimativa	Erro padrão	Estatística	p.value
(Intercepto)	32.885	1.535	21.417	2.54×10^{-19}
x_1	-0.053	0.006	-8.660	1.55×10^{-9}
x_6	0.959	0.670	1.431	0.163

A análise de variância traz a informação de que o modelo num todo é significativo, porém a adição da variável x_6 por si só não é significativa para melhorar o modelo. O teste t também indica que a estimativa para o coeficiente da regressão da variável x_6 não é significativa.

Apenas a variável x_1 e o coeficiente do intercepto são altamente significantes para o modelo.