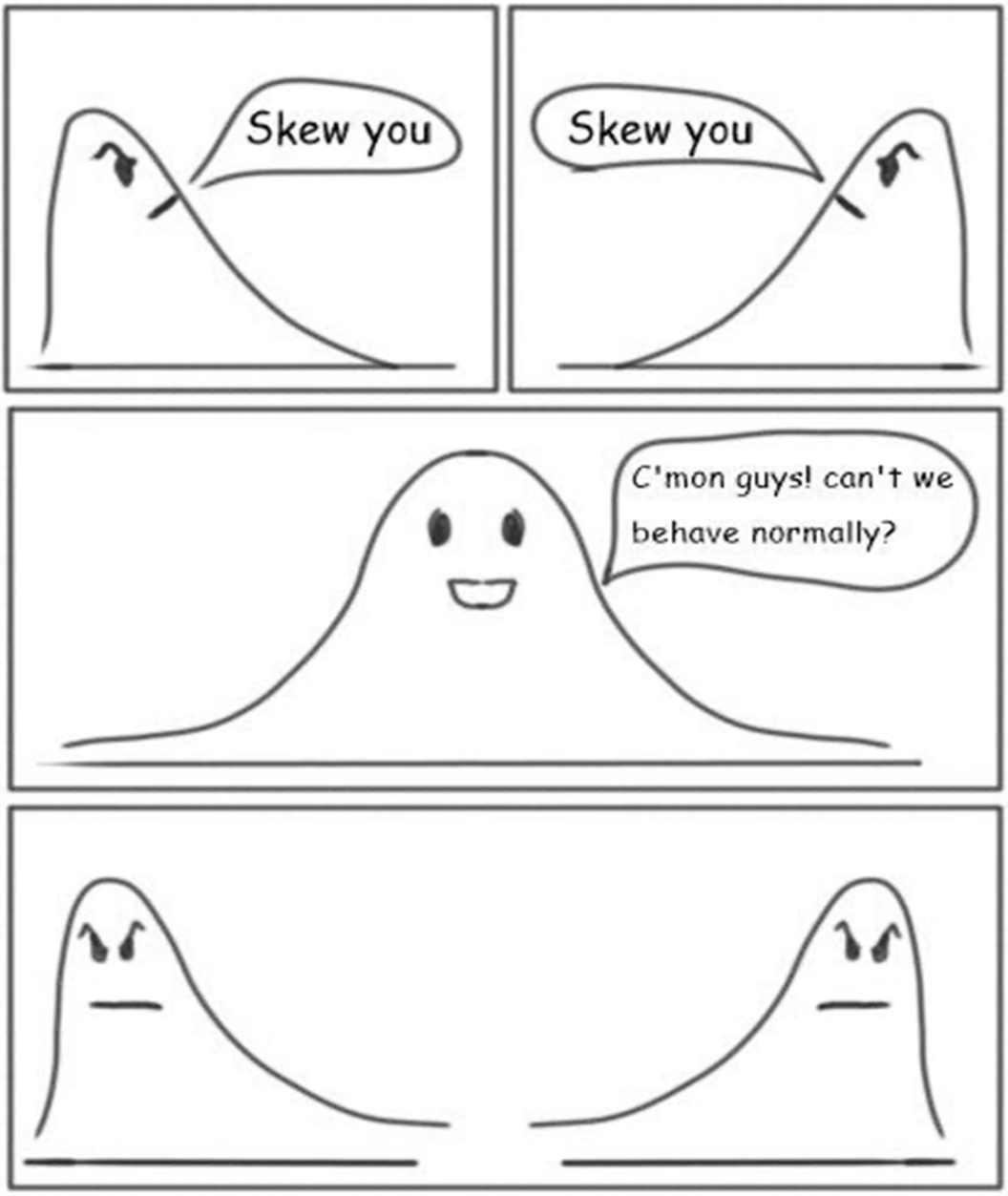


Trabalho R - AEII

Hugo Veríssimo 75695
Mateus Botequilha 75521

19 de dezembro de 2023

UAlg
UNIVERSIDADE DO ALGARVE

Conteúdo

1 Introdução

No cenário atual da análise estatística, a utilização de ferramentas computacionais é fundamental para explorar, analisar e interpretar conjuntos de dados. Neste contexto, a linguagem de programação R destaca-se como uma ferramenta extremamente útil, ao oferecer uma enorme variedade de recursos, de modo a facilitar a realização de análises estatísticas.

Assim sendo, este trabalho tem como intuito explorar a linguagem de programação referida, com foco principal direcionado para a compreensão das funcionalidades específicas do R no contexto de análises estatísticas, em particular, a sua aplicação nas técnicas de análise de variância e regressões lineares múltiplas.

2 Exercício 1

O conjunto de dados "penguins" do package "palmerpenguins" inclui medidas para três espécies de pinguins (Adélie, Chinstrap e Gentoo) da ilha no Arquipélago Palmer, relativas a comprimento das barbatanas, massa corporal, dimensões do bico e sexo. O conjunto de dados contém 8 variáveis para 344 pinguins.

```
library(palmerpenguins)
penguins
```

```
## # A tibble: 344 x 8
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Torgersen           39.1          18.7               181        3750
##  2 Adelie  Torgersen           39.5          17.4               186        3800
##  3 Adelie  Torgersen           40.3          18                 195        3250
##  4 Adelie  Torgersen           NA            NA                  NA          NA
##  5 Adelie  Torgersen           36.7          19.3               193        3450
##  6 Adelie  Torgersen           39.3          20.6               190        3650
##  7 Adelie  Torgersen           38.9          17.8               181        3625
##  8 Adelie  Torgersen           39.2          19.6               195        4675
##  9 Adelie  Torgersen           34.1          18.1               193        3475
## 10 Adelie  Torgersen           42            20.2               190        4250
## # i 334 more rows
## # i 2 more variables: sex <fct>, year <int>
```

No entanto, para este trabalho apenas nos interessam 3 variáveis: espécies (*species*), sexo (*sex*) e a massa corporal do pinguim em gramas (*body_mass_g*).
Um problema que poderemos enfrentar ao analisar o conjunto de dados é o facto do mesmo ter valores em falta (NA) em algumas células, o que pode impactar a precisão e fiabilidade das análises.

2.1 Tratamento dos dados

De modo a simplificar o DataFrame e a dar a volta ao problema referido anteriormente, iremos criar um novo DataFrame (*peng_clean*) apenas com as colunas que iremos utilizar e remover do mesmo as linhas que contêm valores NA.
Adicionalmente, também temos de verificar se as colunas estão prontas para ser utilizadas, isto é, dado que as colunas *species* e *sex* têm valores qualitativos, há que verificar se as mesmas são tratadas como fatores para o R. A coluna *body_mass_g* não terá qualquer problema dado que a mesma apenas contém valores quantitativos.

```
# selecionar as colunas que queremos
peng_clean <- penguins[, c("species", "sex", "body_mass_g")]

# remover linhas com NA
peng_clean <- na.omit(peng_clean)
```

```
# verificar a classe das colunas
cat(class(peng_clean$species), "&", class(peng_clean$sex))
```

```
## factor & factor
```

Como se pode verificar, ambas as colunas são do tipo *factor*, ou seja, são tratadas como fatores, tal como era desejado. Ademais, também já selecionámos as colunas que serão utilizadas e removemos as linhas que continham NA, pelo que já podemos utilizar o nosso novo DataFrame.

```
summary(peng_clean)
```

```
##       species        sex       body_mass_g
##  Adelie   :146   female:165   Min.   :2700
##  Chinstrap: 68   male  :168   1st Qu.:3550
##  Gentoo   :119                Median :4050
##                               Mean   :4207
##                               3rd Qu.:4775
##                               Max.   :6300
```

```
peng_clean
```

```
## # A tibble: 333 x 3
##    species sex    body_mass_g
##    <fct>   <fct>        <int>
##  1 Adelie  male          3750
##  2 Adelie  female        3800
##  3 Adelie  female        3250
##  4 Adelie  female        3450
##  5 Adelie  male          3650
##  6 Adelie  female        3625
##  7 Adelie  male          4675
##  8 Adelie  female        3200
##  9 Adelie  male          3800
## 10 Adelie  male          4400
## # i 323 more rows
```

2.2 Verificação de pressupostos

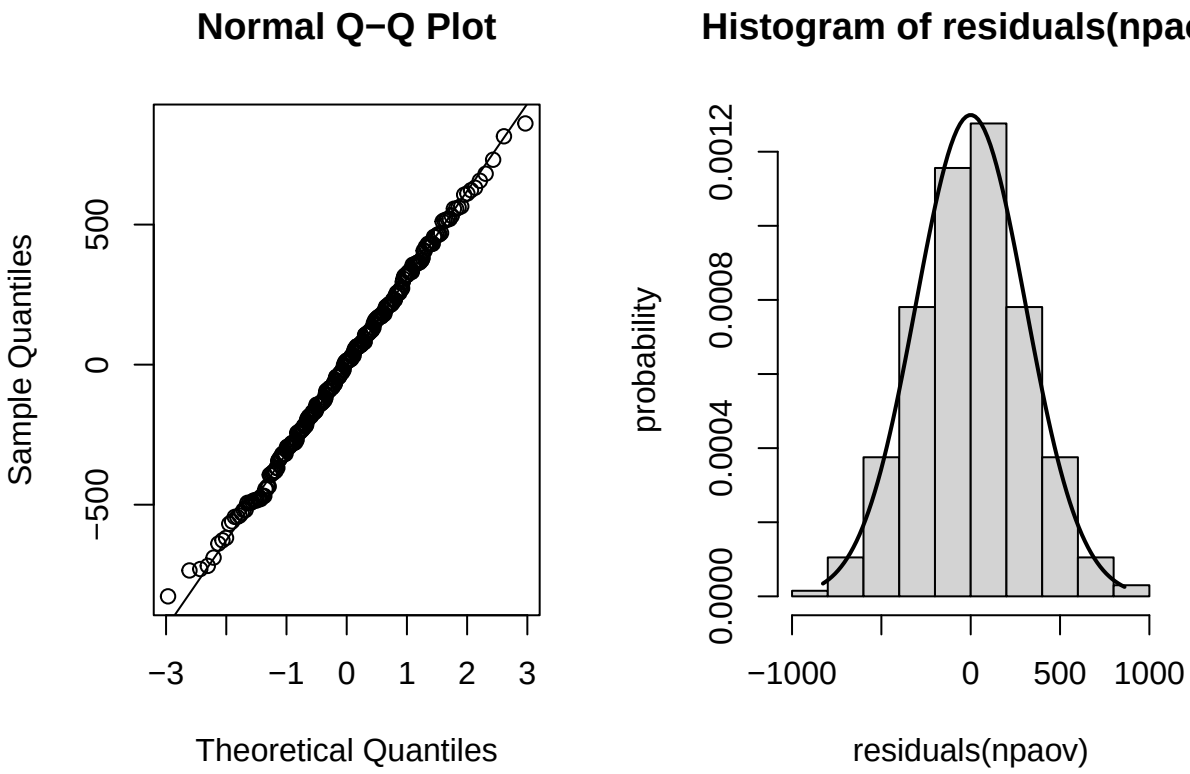
Antes de avançarmos para a análise de variância, é essencial realizar a verificação dos pressupostos da ANOVA. Isto implica analisar a normalidade dos dados e dos resíduos, a independência dos dados e dos resíduos e, ainda, a homogeneidade das variâncias.
A verificação destes pressupostos é fundamental para garantir a validade dos resultados obtidos na análise de variância, de modo a conseguirmos ter uma maior significância nas interpretações. Isto é, após confirmarmos que os mesmos se verificam, estaremos mais confiantes na robustez dos resultados que obtermos a partir da ANOVA.

```r
# anexar o DF ao environment
attach(peng_clean)

# para podermos analisar o fator residuos
npaov <- aov(formula = body_mass_g ~ species * sex, data = peng_clean)
```

2.2.1 Normalidade dos resíduos e das observações

Primeiramente, vamos analisar a normalidade dos resíduos. Contudo, antes de realizarmos o teste formal para testar o pressuposto, decidimos adotar uma abordagem visual para obter uma primeira impressão da distribuição dos resíduos. Com este objetivo, vamos utilizar um gráfico *Quantile-Quantile* para comparar os quantis dos resíduos da amostra com os quantis de uma distribuição normal, e um histograma, de modo a comparar a distribuição dos resíduos da amostra com uma distribuição normal de média 0 e de variância igual à dos referidos.

```r
# mudar o layout grafico para o tipo i,j
par(mfrow = c(1, 2))

# grafico qq
qqnorm(residuals(npaov))
qqline(residuals(npaov))

# histograma
hist(residuals(npaov), probability = TRUE, ylab = "probability")
xfit <- seq(min(residuals(npaov)), max(residuals(npaov)), length=100)
yfit_residuals <- dnorm(xfit, mean=0, sd = sqrt(var(residuals(npaov))))
lines(xfit, yfit_residuals, col = "black", lwd = 2)
```




```
# voltar ao layout grafico normal
par(mfrow = c(1, 1))
```

Como se pode observar em ambos os gráficos, os resíduos da amostra exibem uma grande proximidade em relação às linhas que representam a distribuição normal, pelo que será de esperar que os resíduos sigam uma distribuição normal.
Para corroborar esta observação, de forma a obtermos uma validação mais formal da normalidade dos resíduos, devemos realizar o teste de normalidade de Shapiro-Wilk.

$H_0 : Os\ residuos\ seguem\ uma\ distribuicão\ normal\ vs\ H_1 : Os\ residuos\ não\ seguem\ uma\ distribuicão\ normal$

```
shapiro.test(residuals(npaov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(npaov)
## W = 0.99776, p-value = 0.9367
```

$p-value = 0.9367 > 0.05 = \alpha \Rightarrow$ não rejeitamos H_0 para o nível de significância de 5%, isto é, existe evidência estatística dos resíduos da amostra seguirem uma distribuição normal.

Para além da análise da normalidade dos resíduos, também é importante testar a normalidade das observações. Com este propósito, devemos realizar testes de Shapiro-Wilk entre cada grupo de observações, ou seja, um teste para cada combinação entre *species* e *sex*, dado que são estas as variáveis explicativas.

$$H_0 : O\ grupo_{species,sex}\ segue\ uma\ distribuicão\ normal$$

$$H_1 : O\ grupo_{species,sex}\ não\ segue\ uma\ distribuicão\ normal$$

```
aggregate(body_mass_g ~ species * sex, data = peng_clean,
          function(x) shapiro.test(x)$p.value)
```

```
##     species    sex body_mass_g
## 1    Adelie female   0.1985303
## 2 Chinstrap female   0.3055292
## 3    Gentoo female   0.5106595
## 4    Adelie   male   0.4159824
## 5 Chinstrap   male   0.8910238
## 6    Gentoo   male   0.9850457
```

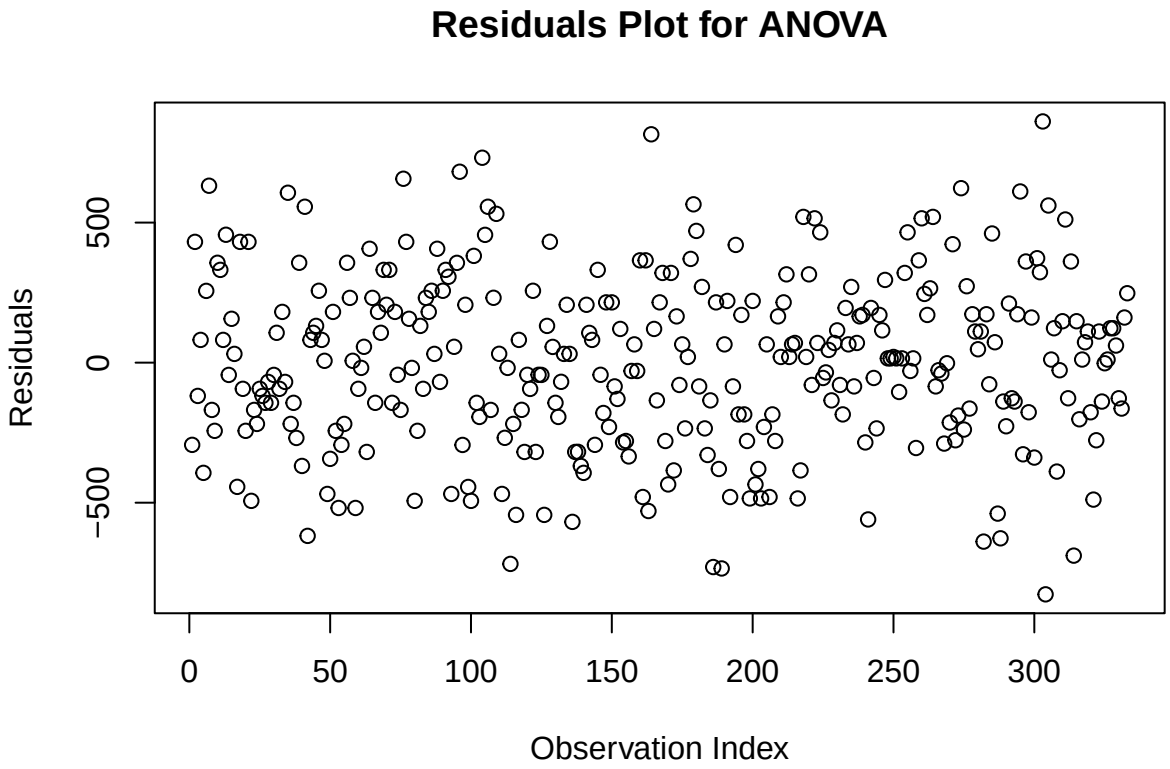
$min\{p - value_{species,sex}\} \approx 0.1985 > 0.05 = \alpha \Rightarrow$ não rejeitamos nenhum dos H_0 para o nível de significância de 5%, isto é, existe evidência estatística de que todos os grupos seguem uma distribuição normal.

Desta forma, após termos analisado os resultados dos testes de Shapiro-Wilk, podemos concluir que tanto os resíduos como as observações seguem distribuições normais, pelo que esta constatação valida a pressuposição de normalidade.

2.2.2 Independência dos resíduos

Após a análise do pressuposto da normalidade, devemos analisar a validade do pressuposto da independência, isto é, devemos agora verificar se os resíduos são independentes entre si. Com este objetivo, vamos analisar a independência resíduos através de uma abordagem visual.

```
plot(residuals(npaov), ylab = "Residuals", xlab = "Observation Index",
     main = "Residuals Plot for ANOVA")
```

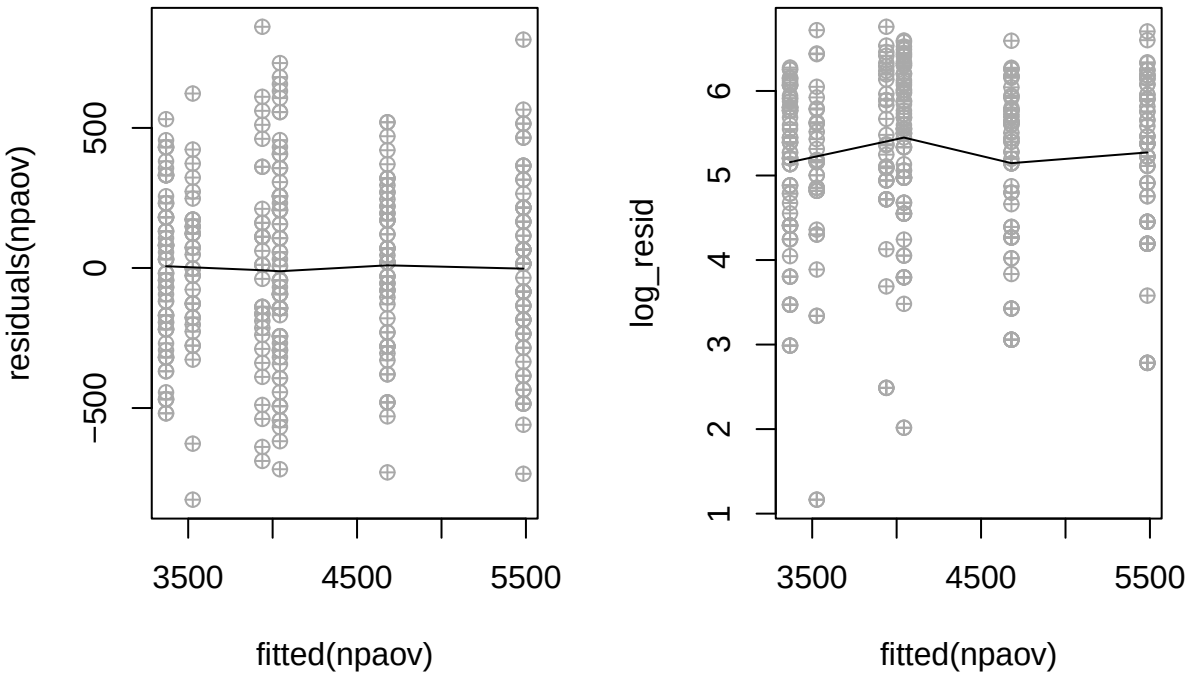


Através da visualização do gráfico dos resíduos, podemos verificar que não existe qualquer padrão entre os mesmos, mas sim uma distribuição aleatória entre eles, ao longo do eixo horizontal, o que nos permite concluir que há independência dos resíduos, ou seja, verifica-se o pressuposto.

2.2.3 Homogeneidade das variâncias

Atendendo ao pressuposto que nos falta verificar, a homogeneidade das variâncias, este pode ser analisado de forma gráfica, através da análise da relação entre os resíduos e os valores ajustados, e também através da realização de um teste estatístico, o teste de Bartlett.
Comecemos pela forma gráfica.

```
par(mfrow = c(1, 2))

# dispersao entre valores ajustados e residuos
plot(fitted(npaov), residuals(npaov), col = "darkgray", pch = 10)
# linha que melhor se ajusta aos padroes dos residuos
lines(lowess(residuals(npaov) ~ fitted(npaov)), col = "black")

# aplicar transformacao logaritmica para tornar padroes mais evidentes
log_resid <- log1p(abs(residuals(npaov)))
plot(fitted(npaov), log_resid, col = "darkgray", pch = 10)
lines(lowess(log_resid ~ fitted(npaov)), col = "black")
```




```
par(mfrow = c(1, 1))
```

Pelo facto da linha que melhor se ajusta à dispersão dos resíduos ser significativamente horizontal, pode-se verificar que a dispersão dos mesmos é constante, pelo que o pressuposto da homogeneidade de variâncias se verifica.
De qualquer forma, analisemos agora o pressuposto, novamente, mas através do teste de Bartlett.

```
detach(peng_clean)
# introduzir coluna nova do tipo "species.sex" (ex.: Gentoo.female)
peng_clean$bart <- interaction(peng_clean$species, peng_clean$sex)
attach(peng_clean)

summary(bart)
```

```
##    Adelie.female Chinstrap.female    Gentoo.female      Adelie.male
##               73                34               58               73
##   Chinstrap.male      Gentoo.male
##               34               61
```

Note-se que todos os grupos têm pelo menos 5 observações, pelo que os resultados serão mais significantes, dada a sensibilidade deste teste ao tamanho das amostras. Avancemos com o teste.

$$H_0: \exists\, homogeneidade\ de\ variâncias\ vs\ H_1: \nexists\, existe\ homogeneidade\ de\ variâncias$$

```
bartlett.test(body_mass_g ~ bart, data = peng_clean)
```

```
##
##  Bartlett test of homogeneity of variances
```

```
##
## data:  body_mass_g by bart
## Bartlett's K-squared = 7.6908, df = 5, p-value = 0.1741
```

$p-value = 0.1741 > 0.05 = \alpha \Rightarrow$ não rejeitamos H_0 para o nível de significância de 5%, isto é, existe evidência estatística de não haverem diferenças significativas entre as variâncias de diferentes grupos. Assim, verifica-se o pressuposto da homogeneidade das variâncias, tal como já havíamos verificado anteriormente.

```
detach(peng_clean)
```

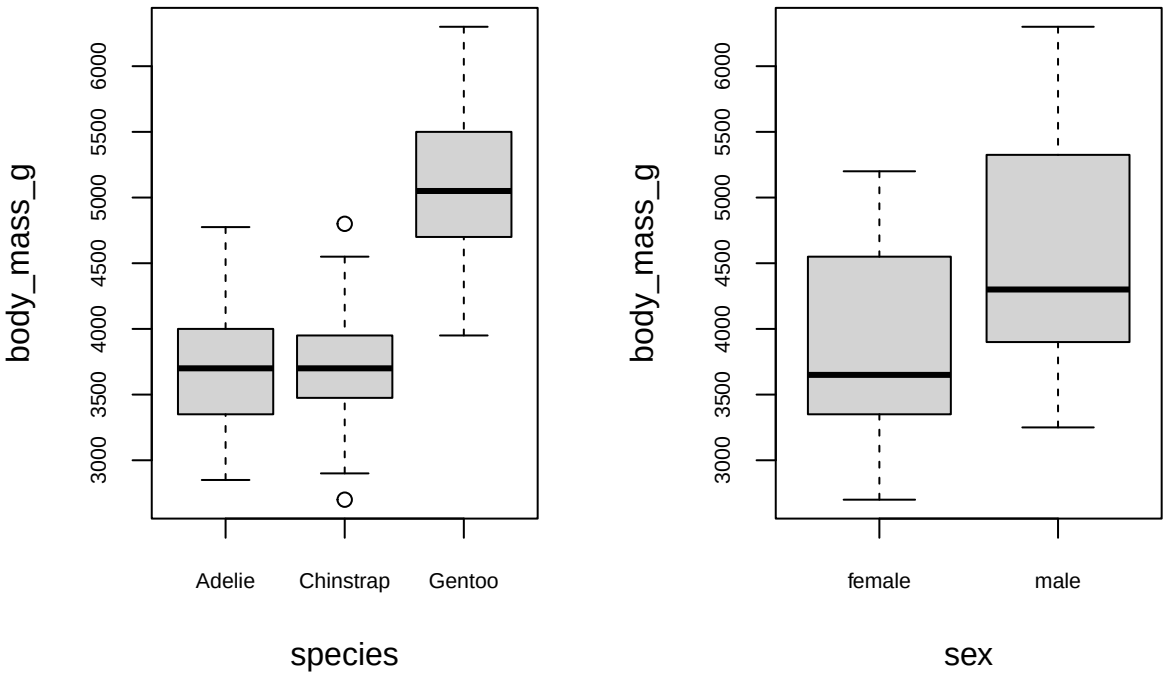
Em suma, tendo por base as análises realizadas, verificamos que os dados atendem aos pressupostos desejados, o que fortalece a validade estatística dos próprios, estabelecendo uma base sólida e confiável para a interpretação dos resultados subsequentes da análise de variância, assegurando a robustez e a precisão das conclusões extraídas a partir desse estudo.

2.3 Análise descritiva dos dados

Realizada a verificação dos pressupostos para o nosso conjunto de dados, estamos prontos para realizar uma análise descritiva. Isto implica calcular as médias por variável (*species* e *sex*) e por grupo formado pela combinação das mesmas (*species.sex*), o que nos irá permitir ter uma perceção dos resultados que devemos esperar ao realizar a análise de variância. Para além do cálculo das médias, também podemos criar gráficos de interação, de modo a observar a presença, ou não, de interação entre os fatores.

Comecemos pelo cálculo das médias por variável, mas ao invés de nos restringirmos a números, exploremos visualmente as mesmas.

```
attach(peng_clean)
par(mfrow = c(1, 2))

# media species e sex, atraves de caixas de bigodes
plot(species, body_mass_g, ylab = "body_mass_g", xlab = "species", cex.axis = 0.7)
plot(sex, body_mass_g, ylab = "body_mass_g", xlab = "sex", cex.axis = 0.7)
```




```
par(mfrow = c(1, 1))
```

Note-se que apenas não há diferenças significativas entre as médias das espécies Adelie e Chinstrap, pelo que será de esperar que rejeitemos a hipótese de haver igualdade entre as médias tanto na variável *species*, como na variável *sex*.

Analisemos agora as médias entre os grupos formados por cada tipo de *species* e *sex*.

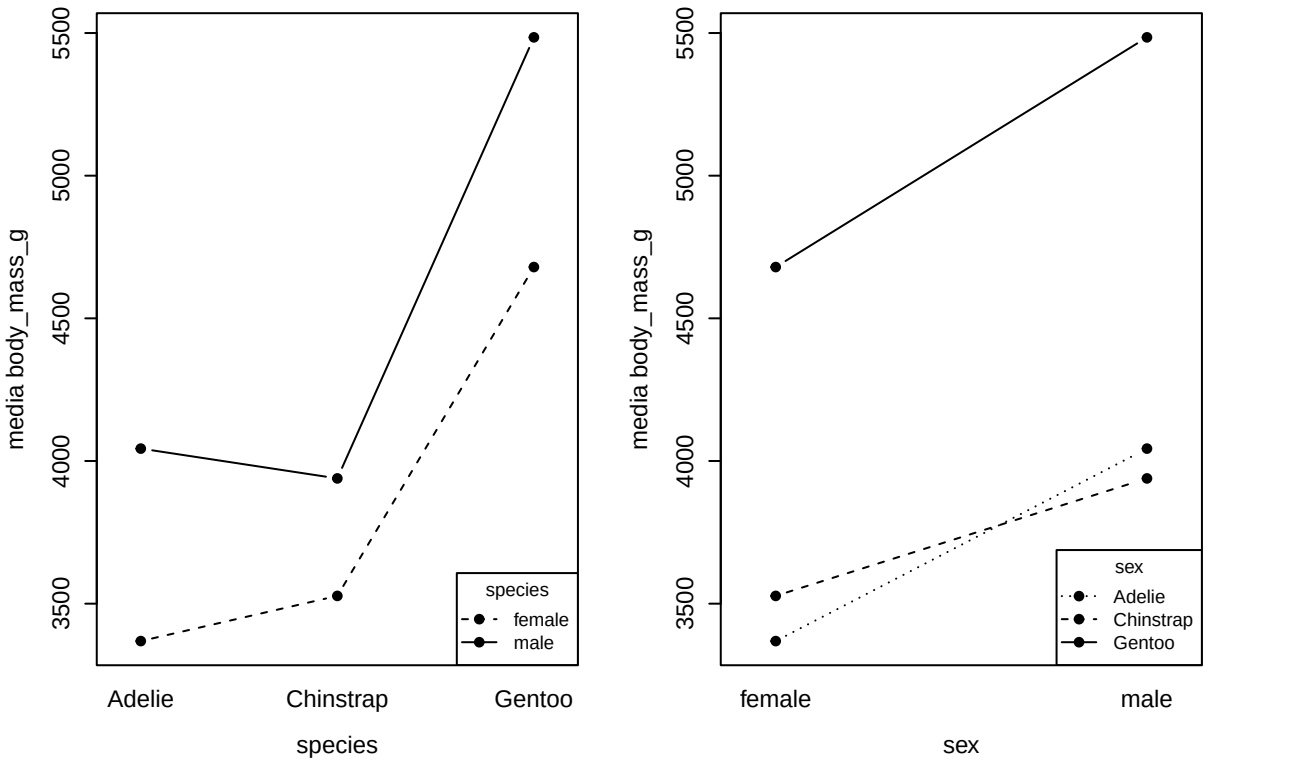
```
# media species.sex
aggregate(body_mass_g ~ species * sex, data = peng_clean, FUN = mean)
```

```
##     species    sex body_mass_g
## 1    Adelie female    3368.836
## 2 Chinstrap female    3527.206
## 3    Gentoo female    4679.741
## 4    Adelie   male    4043.493
## 5 Chinstrap   male    3938.971
## 6    Gentoo   male    5484.836
```

Como podemos verificar, a análise das diferenças entre as médias de forma numérica é mais trabalhosa do que através de representações gráficas. Contudo, podemos, ainda assim, observar que apenas há proximidade nas médias entre as espécies Adelie e Chinstrap, quando ambos são ou machos ou fêmeas.

No que diz respeito à interação entre os fatores, procederemos à criação de gráficos de interação para verificar a existência, ou não, da mesma, tal como referido.

```
par(mfrow = c(1, 2))
# mudar tamanho da fonte (lab, axis, tudo) em %
par(cex.lab = 1.3, cex.axis = 1.3, cex=0.6)

# fazer os graficos de interacao
with(peng_clean, interaction.plot(species, sex, body_mass_g,
                            type = "b", pch = 19, fixed = T,
                            xlab = "species", ylab = "media body_mass_g", legend = FALSE))
legend("bottomright", legend = c("female", "male"),
       title = "species", lty = c(2,1), pch = 19)

with(peng_clean, interaction.plot(sex, species, body_mass_g,
                                  type = "b", pch = 19, fixed = T,
                                  xlab = "sex", ylab = "media body_mass_g", legend = FALSE))
legend("bottomright", legend = c("Adelie", "Chinstrap", "Gentoo"),
       title = "sex", lty = c(3,2,1), pch = 19)
```




```
par(mfrow = c(1, 1))
detach(peng_clean)
```

Como podemos observar, no primeiro gráfico, há falta de paralelismo. O mesmo se passa no segundo gráfico, de forma ainda mais evidente, pelo facto de haver interseção de duas retas. Estas observações indicam-nos que haverá interação entre os fatores.

2.4 Análise de variância (ANOVA)

Finalmente, após o tratamento dos dados, a verificação dos pressupostos e uma breve análise descritiva, podemos realizar a análise de variância (ANOVA). Esta análise permite-nos testar as seguintes hipóteses:

- $H_0' : \mu_{Adelie} = \mu_{Chinstrap} = \mu_{Gentoo} = \mu \; vs \; H_1' : \exists_i : \mu_i \neq \mu$
- $H_0'' : \mu_{female} = \mu_{male} = \mu \; vs \; H_1'' : \exists_j : \mu_j \neq \mu$
- $H_0''' : \nexists \, interação \, entre \, os \, fatores \, species \, e \, sex \; vs \; H_1''' : \exists \, interação \, entre \, os \, fatores \, species \, e \, sex$

```
summary(npaov)
```

```
##              Df    Sum Sq  Mean Sq F value   Pr(>F)
## species       2 145190219 72595110 758.358  < 2e-16 ***
## sex           1  37090262 37090262 387.460  < 2e-16 ***
## species:sex   2   1676557   838278   8.757 0.000197 ***
## Residuals   327  31302628    95727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comecemos por analisar a interação entre os fatores:
$p - value = 0.000197 < 0.05 = \alpha \Rightarrow$ rejeitamos H_0''' para o nível de significância de 5%, isto é, existe evidência estatística de haver interação significativa entre os fatores *species* e *sex*, tal como já havíamos previsto na análise descritiva.

De seguida, analisemos os níveis médios do fator *species*:
$p - value < 2e - 16 < 0.05 = \alpha \Rightarrow$ rejeitamos H_0' para o nível de significância de 5%, isto é, existe evidência estatística de haver diferenças significativas entre o peso médio dos pinguins para os níveis do fator *species*, quando considerados relativamente aos níveis do fator *sex*, em média.

E por último, analisemos os níveis médios do fator *sex*:
$p - value < 2e - 16 < 0.05 = \alpha \Rightarrow$ rejeitamos H_0'' para o nível de significância de 5%, isto é, existe evidência estatística de haver diferenças significativas entre o peso médio dos pinguins para os níveis do fator *sex*, quando considerados relativamente aos níveis do fator *species*, em média.

Pelo facto de termos verificado que existem diferenças entre o peso médio dos pinguins, tanto para os níveis do fator *species*, como para os níveis do fator *sex*, tal como era de esperar, tendo em conta as conclusões tiradas a partir da análise descritiva, temos agora de averiguar que grupos (*species:sex*, dado que os fatores têm interação) têm, ou não, médias significativamente diferentes. Para realizar estas comparações múltiplas iremos utilizar o teste de Tukey, tanto numérica como graficamente.

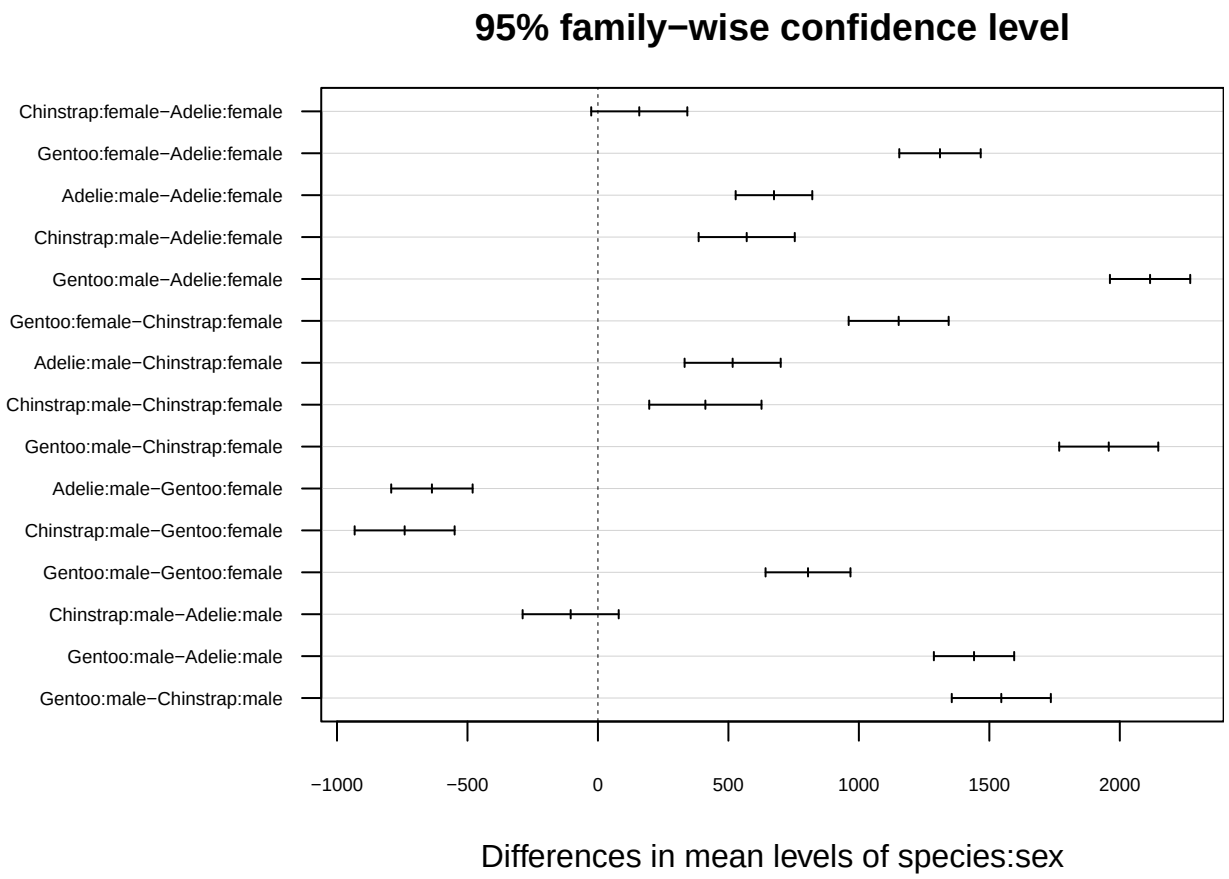
$$H_0 : \mu_{species_i, sex_i} = \mu_{species_j, sex_j} \; vs \; H_1 : \mu_{species_i, sex_i} \neq \mu_{species_j, sex_j}$$

```
TukeyHSD(npaov, "species:sex")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = body_mass_g ~ species * sex, data = peng_clean)
##
## $`species:sex`
##                                       diff       lwr       upr     p adj
## Chinstrap:female-Adelie:female   158.3703  -25.7874  342.5279 0.1376213
## Gentoo:female-Adelie:female     1310.9058 1154.8934 1466.9181 0.0000000
## Adelie:male-Adelie:female        674.6575  527.8486  821.4664 0.0000000
## Chinstrap:male-Adelie:female     570.1350  385.9773  754.2926 0.0000000
## Gentoo:male-Adelie:female       2116.0004 1962.1408 2269.8601 0.0000000
## Gentoo:female-Chinstrap:female  1152.5355  960.9603 1344.1107 0.0000000
## Adelie:male-Chinstrap:female     516.2873  332.1296  700.4449 0.0000000
## Chinstrap:male-Chinstrap:female  411.7647  196.6479  626.8815 0.0000012
## Gentoo:male-Chinstrap:female    1957.6302 1767.8040 2147.4564 0.0000000
## Adelie:male-Gentoo:female       -636.2482 -792.2606 -480.2359 0.0000000
## Chinstrap:male-Gentoo:female    -740.7708 -932.3460 -549.1956 0.0000000
## Gentoo:male-Gentoo:female        805.0947  642.4300  967.7594 0.0000000
## Chinstrap:male-Adelie:male      -104.5226 -288.6802   79.6351 0.5812048
## Gentoo:male-Adelie:male         1441.3429 1287.4832 1595.2026 0.0000000
## Gentoo:male-Chinstrap:male      1545.8655 1356.0392 1735.6917 0.0000000
```

Tendo em conta o teste de Tukey realizado, podemos reparar que apenas as combinações *Chinstrap:female-Adelie:female* e *Chinstrap:male−Adelie:male* têm p-values superiores a 0.05 (0.1376213 e 0.5812048, respetivamente), pelo que as restantes combinações de grupos rejeitam H_0, mas estas não. Isto significa que para um nível de significância de 5%, existe evidência estatística de que os pinguins fêmea das espécies Chinstrap e Adelie não têm diferenças significativas no seu peso médio, tal como os pinguins machos das mesmas espécies, e de que os pesos médios dos restantes grupos de pinguins (*species:sex*) têm todos diferenças significativas entre si.

```
# alterar as margens (baixo, esqueda, cima, direita)
par(mar = c(4, 9, 2, 0))
plot(TukeyHSD(npaov, "species:sex"), cex.axis = 0.6, las = 1)
```



Através da análise do gráfico, chegamos às mesmas conclusões mencionadas anteriormente. As combinações que incluem o valor zero no seu intervalo são aquelas que provam estatisticamente a ausência de diferenças significativas entre os pesos médios entre os grupos de pinguins a serem comparados. Isto é, as combinações que contêm o valor zero no seu intervalo, são aquelas que não rejeitam a hipótese nula (H_0) definida anteriormente, enquanto que as restantes a rejeitam.

Note-se que todas as conclusões derivadas da análise de variância já haviam sido antecipadas durante a análise descritiva, evidenciando, desta forma, concordância entre as previsões e as conclusões, como seria de esperar.

3 Exercício 2

O conjunto de dados "sat" do package "faraway" foi obtido com o objetivo de estudar a relação entre as despesas dos alunos com a educação no ensino público e os resultados obtidos no exame SAT.

```
library(faraway)
head(sat, 12)
```

```
##             expend ratio salary takers verbal math total
## Alabama      4.405  17.2 31.144      8    491  538  1029
## Alaska       8.963  17.6 47.951     47    445  489   934
## Arizona      4.778  19.3 32.175     27    448  496   944
## Arkansas     4.459  17.1 28.934      6    482  523  1005
## California   4.992  24.0 41.078     45    417  485   902
## Colorado     5.443  18.4 34.571     29    462  518   980
## Connecticut  8.817  14.4 50.045     81    431  477   908
## Delaware     7.030  16.6 39.076     68    429  468   897
## Florida      5.718  19.1 32.588     48    420  469   889
## Georgia      5.193  16.3 32.291     65    406  448   854
## Hawaii       6.078  17.9 38.518     57    407  482   889
## Idaho        4.210  19.1 29.783     15    468  511   979
```

O conjunto de dados contém 7 variáveis relativas aos resultados de 50 alunos. No entanto, para a nossa regressão linear múltipla, apenas iremos considerar as variáveis despesas (*expend*), razão média de alunos por professor (*ratio*), ordenado (*salary*), percentagem de alunos elegíveis para fazerem o exame (*takers*) e pontuação média total no SAT (*total*).
Para simplificar o conjunto de dados, podemos criar um novo DataFrame, apenas com as variáveis necessárias.

```
sat_clean <- sat[, c("total", "expend", "ratio", "salary", "takers")]
head(sat_clean, 12)
```

```
##             total expend ratio salary takers
## Alabama      1029  4.405  17.2 31.144      8
## Alaska        934  8.963  17.6 47.951     47
## Arizona       944  4.778  19.3 32.175     27
## Arkansas     1005  4.459  17.1 28.934      6
## California    902  4.992  24.0 41.078     45
## Colorado      980  5.443  18.4 34.571     29
## Connecticut   908  8.817  14.4 50.045     81
## Delaware      897  7.030  16.6 39.076     68
## Florida       889  5.718  19.1 32.588     48
## Georgia       854  5.193  16.3 32.291     65
## Hawaii        889  6.078  17.9 38.518     57
## Idaho         979  4.210  19.1 29.783     15
```

3.1 Estimação do modelo de regressão linear múltipla

Após termos selecionado as variáveis que vamos utilizar, podemos agora construir o nosso modelo de regressão linear múltipla, o que nos irá permitir explorar a relação entre as variáveis explicativas (*expend*, *ratio*, *salary*, *takers*) e a variável dependente (*total*).

```
# criar o modelo de regressao linear
sat_lm <- lm(total ~ expend + ratio + salary + takers, data = sat_clean)
sat_lm
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat_clean)
##
## Coefficients:
## (Intercept)       expend        ratio       salary       takers
##    1045.972        4.463       -3.624        1.638       -2.904
```

Estimado o modelo, devemos analisar os coeficientes, ou seja, vamos verificar o impacto de cada variável explicativa na variável dependente e analisar o que significa esse impacto no contexto dos nossos dados.

- β_0 : Estima-se que se todas as variáveis explicativas (*expend*, *ratio*, *salary* e *takers*) forem nulas, então a pontuação média total no SAT (*total*) será de 1045.9715 unidades.
- β_{expend} : Estima-se que por cada variação unitária nas despesas (*expend*), a pontuação média total no SAT (*total*) varie 4.4626 unidades, assumindo tudo o resto constante.
- β_{ratio} : Estima-se que por cada variação unitária na razão média de alunos por professor (*ratio*), a pontuação média total no SAT (*total*) varie -3.6242 unidades, assumindo tudo o resto constante.
- β_{salary} : Estima-se que por cada variação unitária no ordenado (*salary*), a pontuação média total no SAT (*total*) varie 1.6379 unidades, assumindo tudo o resto constante.
- β_{takers} : Estima-se que por cada variação unitária na percentagem de alunos elegíveis para fazerem o exame (*takers*), a pontuação média total no SAT (*total*) varie -2.9045 unidades, assumindo tudo o resto constante.

3.2 Interpretação do modelo estimado

Tendo em conta o modelo de regressão estimado, através da análise do mesmo, para além de podermos interpretar os coeficientes, tal como fizemos, podemos interpretar outros valores, o que nos irá permitir tirar outras conlusões sobre o modelo.

```
summary(sat_lm)
```

```
##
## Call:
## lm(formula = total ~ expend + ratio + salary + takers, data = sat_clean)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## ratio         -3.6242     3.2154  -1.127    0.266
## salary         1.6379     2.3872   0.686    0.496
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

Comecemos por analisar os resultados presentes sobre os testes de significância individuais:

$$H_0 : \beta_i = 0 \, vs \, H_1 : \beta_i \neq 0$$

- $\beta_{expend} : p - value = 0.674 > 0.05 = \alpha \Rightarrow$ não rejeitamos H_0 para o nível de significância de 5%, isto é, existe envidência estatística de que a variável *expend* não é significativa para o modelo que inclui as variáveis *ratio*, *salary* e *takers*.
- $\beta_{ratio} : p - value = 0.266 > 0.05 = \alpha \Rightarrow$ não rejeitamos H_0 para o nível de significância de 5%, isto é, existe envidência estatística de que a variável *ratio* não é significativa para o modelo que inclui as variáveis *expend*, *salary* e *takers*.
- $\beta_{salary} : p - value = 0.496 > 0.05 = \alpha \Rightarrow$ não rejeitamos H_0 para o nível de significância de 5%, isto é, existe envidência estatística de que a variável *salary* não é significativa para o modelo que inclui as variáveis *expend*, *ratio* e *takers*.
- $\beta_{takers} : p - value = 2.61e-16 < 0.05 = \alpha \Rightarrow$ rejeitamos H_0 para o nível de significância de 5%, isto é, existe envidência estatística de que a variável *takers* é significativa para o modelo que inclui as variáveis *expend*, *ratio* e *salary*.

De seguida, podemos verificar que *Adjusted R-squared: 0.809*, o que nos indica que aproximadamente 81% da variação da pontuação média total no SAT (*total*), pode ser explicada pelo modelo estimado.

Por último, também é de elevada importância avaliar a significância do modelo de regressão linear múltipla estimado:

$$H_0 : \beta_0 = \beta_{expend} = \beta_{ratio} = \beta_{salary} = \beta_{takers} = 0 \, vs \, H_1 : \exists_i : \beta_i \neq 0$$

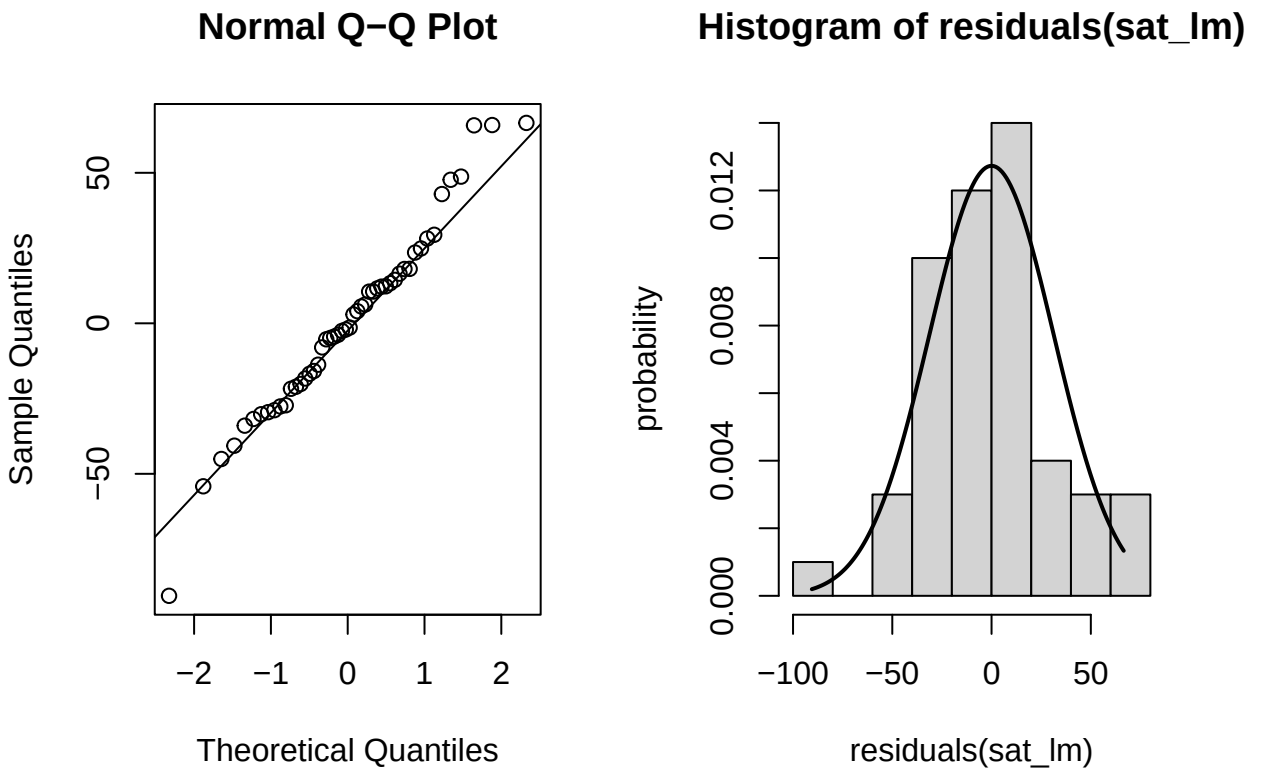
$p - value < 2.2e - 16 < 0.05 = \alpha \Rightarrow$ rejeitamos H_0 para o nível de significância de 5%, isto é, existe evidência estatística de que o modelo ajustado é significativo, ou seja, pelo menos uma das variáveis explicativas tem um efeito significativo sobre a variável dependente.

3.3 Análise de resíduos

Antes de mais, devemos avaliar determinadas suposições sobre o nosso conjunto de dados, ver se as mesmas se verificam. Se as suposições da regressão se mantiverem, os resíduos deverão ser normalmente distribuídos, com valor médio zero e variância constante e independentes entre si.

Comecemos por verificar se são normalmente distribuídos:

```
par(mfrow = c(1,2))

# grafico qq
qqnorm(residuals(sat_lm))
qqline(residuals(sat_lm))

# histograma
hist(residuals(sat_lm), probability = TRUE, ylab = "probability")
xfit <- seq(min(residuals(sat_lm)), max(residuals(sat_lm)), length = 150)
yfit_residuals <- dnorm(xfit, mean = 0, sd = sqrt(var(residuals(sat_lm))))
lines(xfit, yfit_residuals, col = "black",lwd = 2)
```



$H_0 : Os\ residuos\ seguem\ uma\ distribuicão\ normal\ vs\ H_1 : Os\ residuos\ não\ seguem\ uma\ distribuicão\ normal$

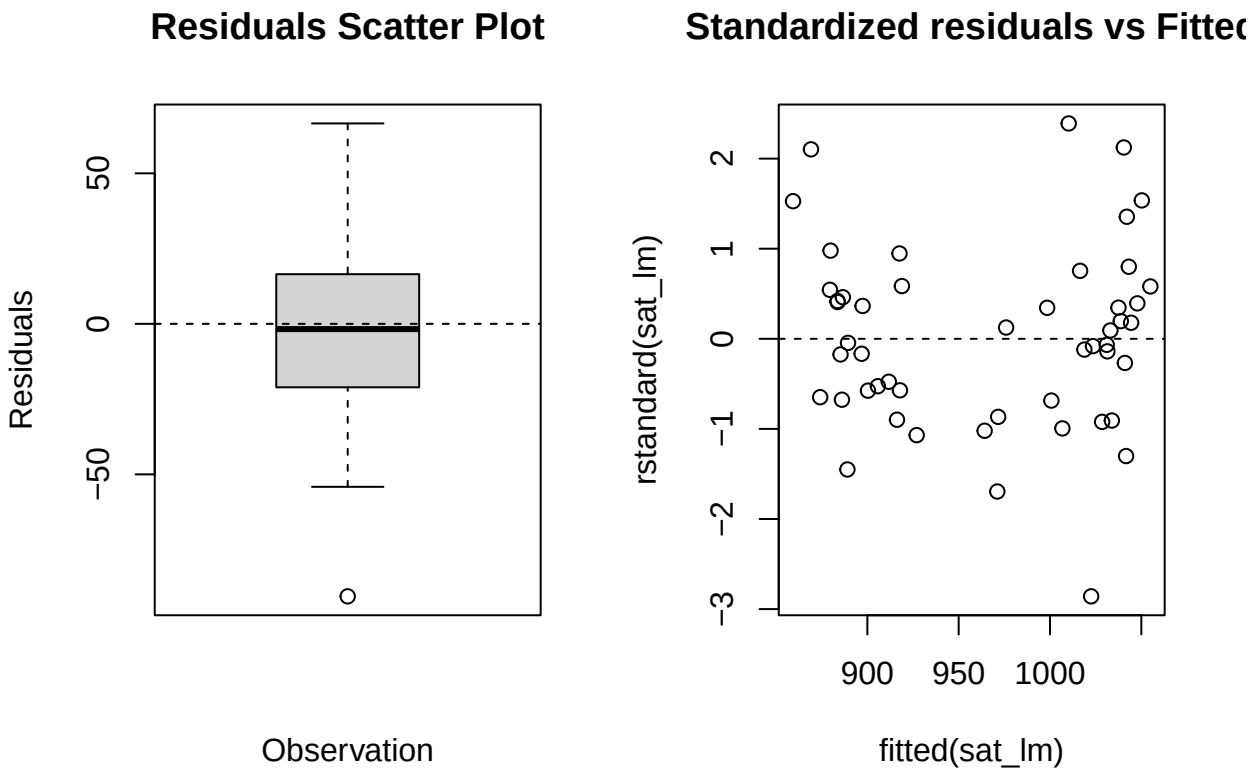
```
par(mfrow = c(1,1))

# teste de Shapiro-Wilk
shapiro.test(residuals(sat_lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(sat_lm)
## W = 0.97691, p-value = 0.4304
```

$p-value = 0.4304 > 0.05 = \alpha \Rightarrow$ não rejeitamos H_0 para o nível de significância de 5%, isto é, existe evidência estatística de que os resíduos seguem uma distribuição normal.
Tendo em conta as representações gráficas, as mesmas reforçam a não rejeição de H_0, pelo facto dos resíduos exibirem uma grande proximidade às linhas que representam a distribuição normal, em cada um dos gráficos.

Verifiquemos se o valor médio é zero e a variância constante:

```
par(mfrow = c(1,2))

# valor medio
boxplot(residuals(sat_lm), main = "Residuals Scatter Plot",
     xlab = "Observation", ylab = "Residuals")
abline(h = 0, col = "black", lty = 2)

# homogeneidade de variancia
plot(fitted(sat_lm), rstandard(sat_lm),
     main = "Standardized residuals vs Fitted")
abline(h = 0, col = "black", lty = 2)
```




```
par(mfrow = c(1,1))
```

Repare-se que no primeiro gráfico podemos observar que a média dos resíduos está muito próxima de zero, pelo que podemos afirmar que o valor médio é zero. Para além disso, se recuarmos aos testes da normalidade, podemos verificar no histograma que a linha com que os resíduos têm uma grande proximidade representa uma distribuição normal de valor médio zero, pelo que essa verificação já estava prevista.
Atendendo ao segundo gráfico, pelo facto dos pontos aparentarem estar distribuídos de forma aleatória, em redor da linha horizonal que representa o zero, temos a verificação de que estamos perante uma variância constante.

Analisemos, por fim, a independência:

```
# carregar os pacotes, sem aviso de conflito
library(zoo, warn.conflicts = FALSE)
library(lmtest)
```

$$H_0 : \nexists\, autocorrelacão\, nos\, residuos\, vs\, H_1 : \exists\, autocorrelacão\, nos\, residuos$$

```
# teste de Durbin-Watson
dwtest(sat_lm)
```

```
##
##  Durbin-Watson test
##
```

```
## data:  sat_lm
## DW = 2.4525, p-value = 0.9459
## alternative hypothesis: true autocorrelation is greater than 0
```

$p-value = 0.9459 > 0.05 = \alpha \Rightarrow$ não rejeitamos H_0 para o nível de significância de 5%, isto é, existe evidência estatística de que não há autocorrelação nos resíduos, ou seja, há independência entre eles.

Podemos assim concluir, através das representações gráficas e dos testes estatísticos realizados, que as suposições da regressão linear múltipla se mantêm, pelo que o modelo parece ser apropriado para explicar a relação entre as variáveis consideradas.

3.4 Otimização do modelo: regressão stepwise

Contudo, nada nos indica que o modelo que estimámos seja o "melhor" para explicar a pontuação média total no SAT (*total*). Para testar se existe um modelo "melhor", ou seja, um modelo que equilibre mais adequadamente a sua complexidade e a explicação da variabilidade da variável dependente, vamos utilizar a função *step* (*stepwise*).
Ademais, devemos notar que a regressão *stepwise*, dependendo do modelo inicial e da direção escolhida, pode diferir. Assim sendo, devemos partir de dois modelos diferentes e comparar as regressões obtidas, pelo que optámos por partir de um modelo explicado por todas as nossas variáveis ($expend + ratio + salary + takers$) e de um modelo sem variáveis explicativas (1). Observemos o que acontece.

```r
# definir os modelos
min.model <- lm(total ~ 1, data = sat_clean)
max.model <- sat_lm

# comeca com tudo mas pode tirar e meter
sat_lm_step.max <- step(max.model, direction = "both")
```

```
## Start:  AIC=353.48
## total ~ expend + ratio + salary + takers
##
##           Df Sum of Sq    RSS    AIC
## - expend  1       191  48315 351.67
## - salary  1       503  48627 352.00
## - ratio   1      1359  49483 352.87
## <none>                 48124 353.48
## - takers  1    168688 216812 426.74
##
## Step:  AIC=351.67
## total ~ ratio + salary + takers
##
##           Df Sum of Sq    RSS    AIC
## <none>                 48315 351.67
```

```
## + expend  1       191  48124 353.48
## - ratio   1      5023  53338 354.62
## - salary  1      6782  55097 356.24
## - takers  1    171126 219441 425.34
```

```
# comeca sem nada mas pode meter e tirar
sat_lm_step.min <- step(min.model, direction = "both", scope = formula(max.model))
```

```
## Start:  AIC=432.5
## total ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + takers  1    215875  58433 357.18
## + salary  1     53078 221230 423.75
## + expend  1     39722 234586 426.68
## <none>                274308 432.50
## + ratio   1      1811 272497 434.17
##
## Step:  AIC=357.18
## total ~ takers
##
##           Df Sum of Sq    RSS    AIC
## + expend  1      8913  49520 350.91
## + salary  1      5095  53338 354.62
## + ratio   1      3336  55097 356.24
## <none>                58433 357.18
## - takers  1    215875 274308 432.50
##
## Step:  AIC=350.91
## total ~ takers + expend
##
##           Df Sum of Sq    RSS    AIC
## <none>                49520 350.91
## + ratio   1       893  48627 352.00
## + salary  1        38  49483 352.87
## - expend  1      8913  58433 357.18
## - takers  1    185066 234586 426.68
```

Ao observarmos os resultados das regressões *stepwise*, verificamos que a escolha do modelo inicial teve impacto no modelo final obtido. Por conseguinte, vamos criar um DataFrame para melhor comparar os dois modelos estimados ao nosso modelo anterior.

```
sat.lm<- c(extractAIC(sat_lm),
              summary(sat_lm)$adj.r.squared)
step.max <- c(extractAIC(sat_lm_step.max),
              summary(sat_lm_step.max)$adj.r.squared)
step.min <- c(extractAIC(sat_lm_step.min),
              summary(sat_lm_step.min)$adj.r.squared)
step_comparacao <- data.frame(sat.lm, step.max, step.min)
rownames(step_comparacao) <- c("Parameters","AIC","Adj. R^2")
step_comparacao
```

```
##                 sat.lm    step.max    step.min
## Parameters   5.0000000   4.0000000   3.0000000
## AIC        353.4755564 351.6740977 350.9055047
## Adj. R^2     0.8089679   0.8123772   0.8117906
```

Através da análise do DataFrame criado, podemos verificar que, em relação ao modelo anterior, os dois novos modelos estimados têm menos parâmetros a estimar, um menor AIC e uma melhor explicação sobre a variação da variável *total*.
Comparando os dois novos modelos, verificamos que o modelo *step.min*, para além de ter menos parâmetros a estimar e um menor AIC, explica aproximadamente 81.2% da variação da variável *total*, tal como o modelo *step.max*.
Desta forma, podemos concluir que o modelo mais adequado para explicar a pontuação média total no SAT (*total*) é o seguinte:

```
sat_lm.step <- sat_lm_step.min
sat_lm.step
```

```
##
## Call:
## lm(formula = total ~ takers + expend, data = sat_clean)
##
## Coefficients:
## (Intercept)       takers       expend
##     993.832       -2.851       12.287
```

3.5 Confirmação do modelo ótimo: testes F-parciais

Uma vez obtido o novo modelo, através do método *stepwise*, podemos confirmar as escolhas das variáveis incluídas no mesmo, através de testes F-parciais. A realização de testes F-parciais vai-nos permitir avaliar se a inclusão de variáveis específicas no modelo melhora significativamente a explicação da variabilidade da variável dependente, ou não.
De modo a facilitar a realização dos testes F-parciais, vamos criar uma função que os realize entre um modelo inicial e vários modelos com apenas mais uma variável que o inicial e, de seguida, apresente o p-value associado a cada um dos testes.

```
testes_f_parcial <- function(formula_0, variaveis) {
  # modelo inicial
  modelo_0 <- lm(as.formula(formula_0), data = sat_clean)

  # fazer anova com a adicao de cada variavel
  resultados_anova <- lapply(variaveis, function(x) {
    formula_i <- paste(formula_0, "+", x)
    modelo_i <- lm(as.formula(formula_i), data = sat_clean)
    anova(modelo_0, modelo_i)})

  # selecionar os p-values
  p_values <- sapply(resultados_anova, function(resultado) resultado$"Pr(>F)"[2])

  # criar dataframe para melhorar a visualizacao
  df <- t(p_values)
  colnames(df) <- variaveis
  rownames(df) <- c("p-values")

  return(df)
}
```

Note-se que os testes têm as seguintes hipóteses associadas:

$$H_0: \text{ A variável } x_i \text{ não é significativa para o modelo vs } H_1: \text{ A variável } x_i \text{ é significativa para o modelo}$$

Comecemos com um modelo inicial sem variáveis explicativas e comparemos com os modelos que contêm cada uma destas.

```
testes_f_parcial("total ~ 1", c("expend", "salary", "ratio", "takers"))
```

```
##               expend     salary     ratio       takers
## p-values 0.006407965 0.00139131 0.5748329 9.791875e-18
```

Como podemos observar, os p-values associados às variáveis *expend*, *salary* e *takers* são menores que 5%, o nosso nível de significância, pelo que rejeitamos os H_0 associados a estas variáveis. Isto é, existe evidência estatística de que cada uma destas variáveis é significativa para o seu modelo.
Contudo, como o menor p-value é o associado à variável *takers*, vamos realizar novamente os testes F-parciais, mas desta vez adicionando a variável *takers* ao modelo incial e comparemos com os modelos idênticos a este, mas com a adição de cada uma das restantes variáveis.

```
testes_f_parcial("total ~ takers", c("expend", "salary", "ratio"))
```

```
##               expend     salary      ratio
## p-values 0.005529458 0.03942052 0.09823538
```

Analogamente, adicionemos agora a variável *expend*.

```
testes_f_parcial("total ~ takers + expend", c("salary", "ratio"))
```

```
##             salary     ratio
## p-values 0.8526704 0.3629065
```

$min\{p - values\} \approx 0.3629 > 0.05 = \alpha \Rightarrow$ não rejeitamos nenhum dos H_0 para o nível de significância de 5%, isto é, existe evidência estatística de que o modelo mais simples, ou seja, aquele que só tem como variáveis explicativas as variáveis *takers* e *expend*, é suficiente para explicar a variável dependente (*total*).
Com isto, podemos reparar que o "melhor" modelo obtido, tendo por base os critérios utilizados nestes testes F-parciais, é idêntico ao "melhor" modelo obtido através do método *stepwise*, o que reforça a validade e consistência do modelo estimado.

3.6 Análise do modelo ótimo

Por fim, analisemos o modelo de regressão linear múltipla que obtemos através da otimização do modelo inicial.

Comecemos por analisar as variáveis do modelo:

```
summary(sat_lm.step)$call
```

```
## lm(formula = total ~ takers + expend, data = sat_clean)
```

```
summary(sat_lm.step)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 993.831659 21.8332335  45.519215 1.579349e-40
## takers       -2.850929  0.2151123 -13.253212 1.729825e-17
## expend       12.286518  4.2243159   2.908523 5.529458e-03
```

Podemos observar que o modelo, como vimos anteriormente, tem como variáveis explicativas as despesas (*expend*) e a percentagem de alunos elegíveis para fazerem o exame (*takers*) e como variável dependente a pontuação média total no SAT (*total*).
Para além disso, podemos observar que os coeficientes β_0, β_{takers} e β_{expend} são agora, aproximadamente, 993.83, -2.85 e 12.29, respetivamente, e que $max\{p-value\} \approx 5.53e-03 < 0.05 = \alpha \Rightarrow$ para um nível de significância de 5%, existe evidência estatística de que cada uma das variáveis é significativa para o modelo.

Verifiquemos agora a significância do modelo através do cálculo do p-value, ao invés da sua observação direta, com o objetivo de explorar mais funções do R:

$$H_0 : \beta_0 = \beta_{takers} = \beta_{expend} = 0 \, vs \, H_1 : \exists_i : \beta_i \neq 0$$

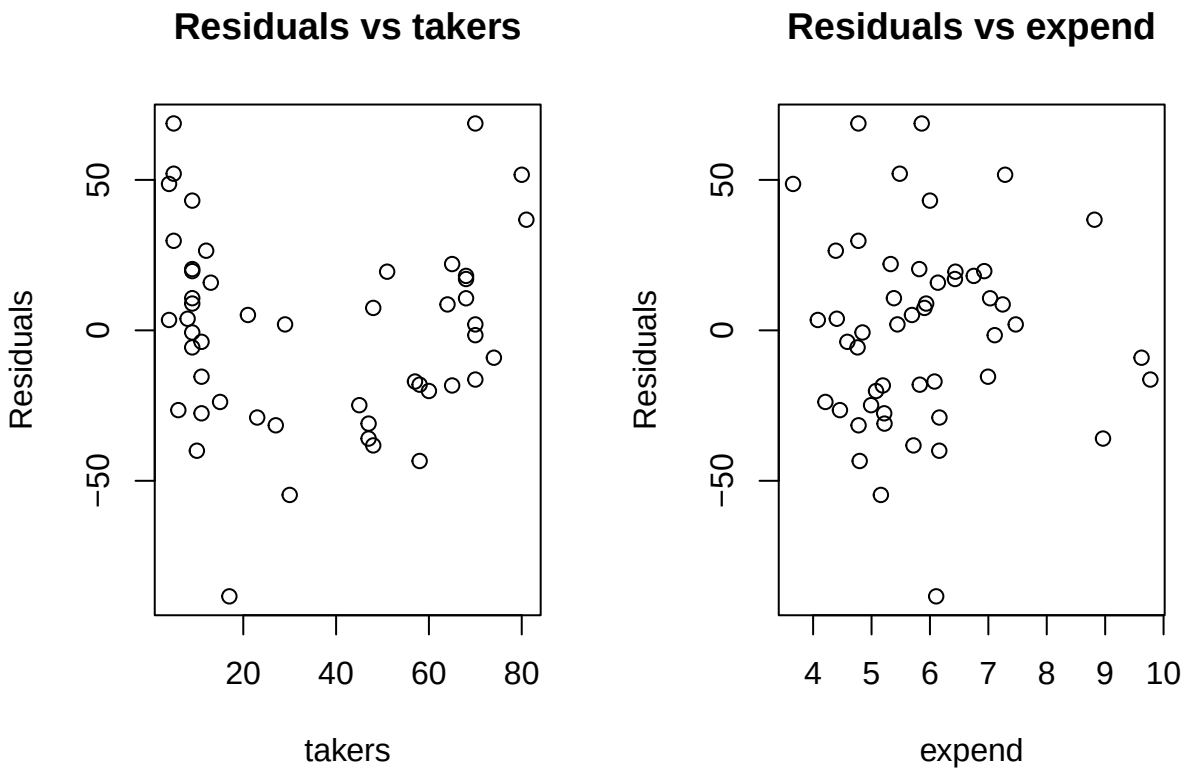
```r
fstatistic_value <- summary(sat_lm.step)$fstatistic[1] # = 106.7
fstatistic_numdf <- summary(sat_lm.step)$fstatistic[2] # = 2
fstatistic_dendf <- summary(sat_lm.step)$fstatistic[3] # = 47

pf(fstatistic_value, fstatistic_numdf, fstatistic_dendf, lower.tail = FALSE)
```

```
##        value
## 3.378819e-18
```

$p - value = 3.378819e - 18 < 0.05 = \alpha \Rightarrow$ rejeitamos H_0 para o nível de significância de 5%, isto é, existe evidência estatística de que o modelo ajustado é significativo, ou seja, pelo menos uma das variáveis explicativas tem um efeito significativo sobre a variável dependente.

De seguida, criemos gráficos de resíduos *vs* fatores, com o objetivo de observar se existem, ou não, padrões na dispersão dos resíduos, consoante a variável explicativa.

```r
par(mfrow = c(1, 2))

# residuos vs takers
plot(sat_clean$takers, residuals(sat_lm.step),
     xlab = "takers", ylab = "Residuals",
     main = "Residuals vs takers")

# residuos vs expend
plot(sat_clean$expend, residuals(sat_lm.step),
     xlab = "expend", ylab = "Residuals",
     main = "Residuals vs expend")
```




```
par(mfrow = c(1, 1))
```

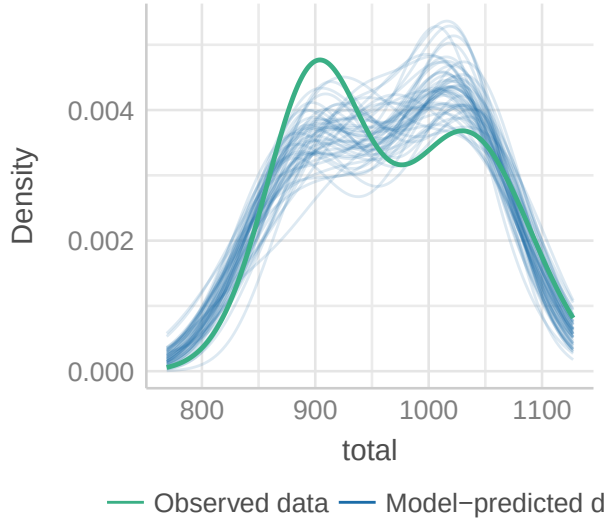
Verifica-se a inexistência de padrões na distribuição dos resíduos, o que nos sugere que o modelo consegue explicar as variações da variável dependente de forma consistente, ao longo dos diferentes níveis dos fatores.

Relembremos que é de elevada importância a verificação de certas condições para que o modelo seja aplicável e que os seus resultados possam ser interpretados de maneira significativa. Verifiquemos as seguintes condições:

```
library(performance, warn.conflicts = FALSE)
check_model(sat_lm.step)
```

Posterior Predictive Check

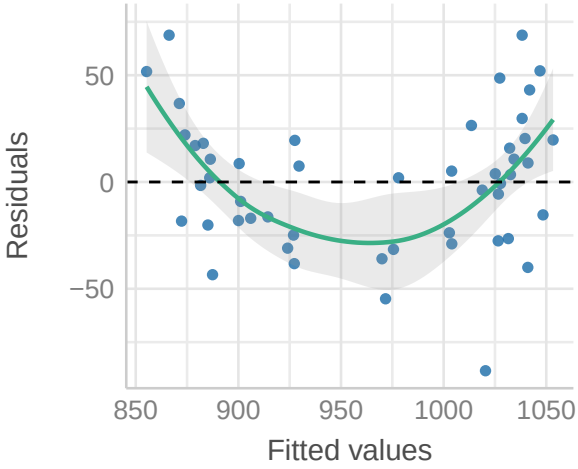
Model−predicted lines should resemble observed da



Observed data Model−predicted d

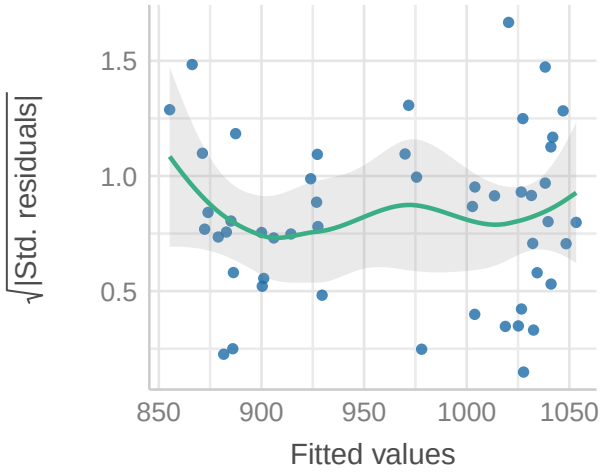
Linearity

Reference line should be flat and horizontal



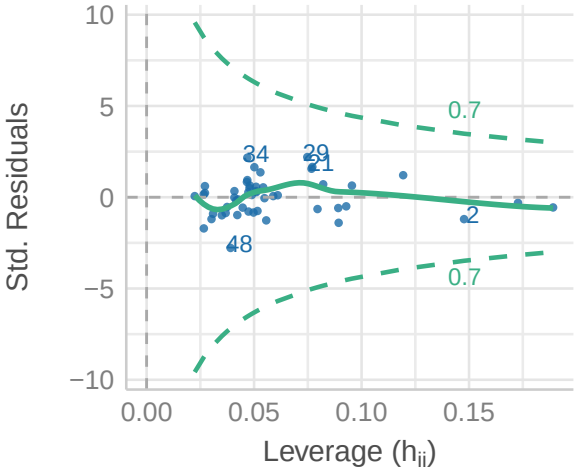
Homogeneity of Variance

Reference line should be flat and horizontal



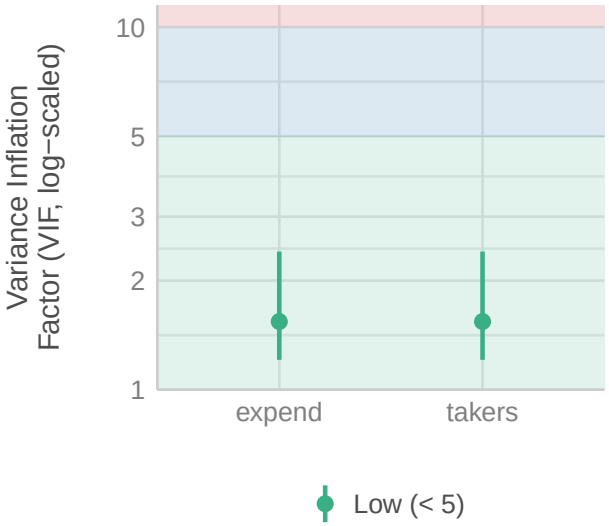
Influential Observations

Points should be inside the contour lines



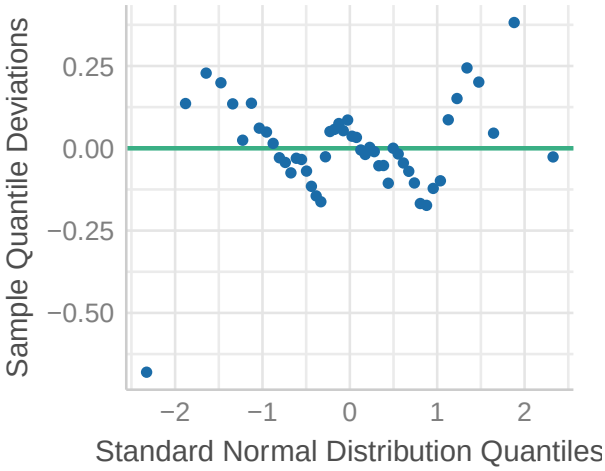
Collinearity

High collinearity (VIF) may inflate parameter uncerta



Normality of Residuals

Dots should fall along the line

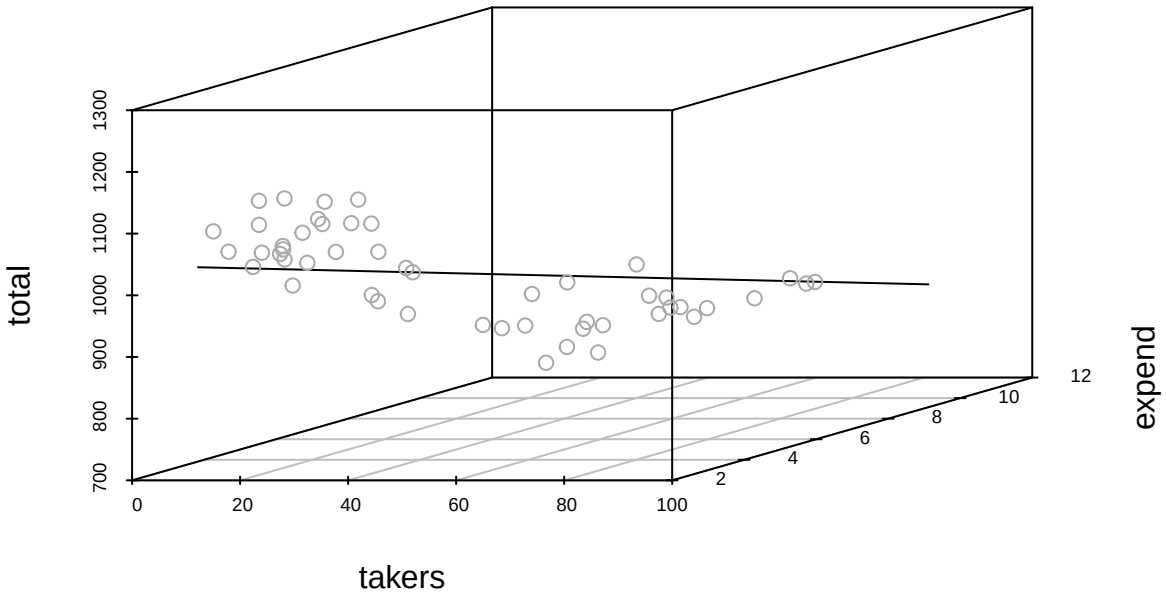


Posterior Predictive Check: podemos verificar que não existem discrepâncias significativas entre os dados reais e os simulados, pelo que podemos afirmar que o modelo se ajusta bem aos dados.
Linearity: devemos notar que a linha que melhor se ajusta aos pontos não é completamente horizontal, mas também não se assemelha significativamente a um "U". Isto indica-nos que a relação entre as variáveis pode não ser linear.
Homogeneity of Variance: notemos que a linha é aproximadamente horizontal, pelo que será expectável que haja homogeneidade de variâncias.
Influential Observations: verifiquemos que os pontos estão todos dentro das linhas a tracejado, pelo que as distâncias de Cook (analisam o impacto de cada ponto de dados na estimação dos coeficientes do modelo de regressão, o que acaba por ser uma verificação de outliers) são respeitadas.
Collinearity: podemos verificar que o VIF é baixo para ambas as variáveis, o que nos indica que não há correlação significativa entre elas.
Normality of Residuals: notemos que os pontos, de forma significativa, alinham-se ao longo da linha horizontal, o que nos permite afirmar que os resíduos têm uma distribuição normal.

Para finalizar, vamos criar um gráfico 3D que compare os valores estimados pelo modelo de regressão linear estimado, face aos valores reais do conjunto de dados.

```
library(scatterplot3d)

x_min <- min(sat_clean$takers) * 0.9
x_max <- max(sat_clean$takers) * 1.1
y_min <- min(sat_clean$expend) * 0.9
y_max <- max(sat_clean$expend) * 1.1

predictions <- predict(sat_lm.step,
                       newdata = data.frame(takers = c(x_min, x_max),
                                            expend = c(y_min, y_max)))

sat_lm.graph <- scatterplot3d(c(x_min,x_max), c(y_min,y_max), predictions,
                      color = "black", type = "l", angle = 30,
                      xlab = "takers", ylab = "expend", zlab = "total",
                      zlim = c(min(sat_clean$total)*0.9, max(sat_clean$total)*1.1),
                      main = "Regressão sat_lm.step", cex.axis = 0.6)

sat_lm.graph$points3d(sat_clean$takers, sat_clean$expend, sat_clean$total,
             col = "darkgray", type = "p")
```

Regressão sat_lm.step



4 Conclusão

A realização deste trabalho deu-nos a oportunidade de explorar mais aprofundadamente as linguagens Rmarkdown e R, tal como o software RStudio. Ademais, através da utlização do R, conseguimos percecionar, em primeira mão, a grande assistência que o mesmo oferece durante a análise de conjuntos de dados, revelando-se particularmente crucial no contexo estatístico.

No âmbito específico da análise de variância, o R destaca-se pela sua capacidade de realizar uma avaliação abrangente das diferenças entre grupos. Além de identificar variações significativas entre médias, o R disponibiliza ferramentas integradas para a verificação de pressupostos essencias, tanto de forma numérica como visual, reforçando a validação dos resultados obtidos na análise de variância.

Quanto à regressão linear múltipla, a linguagem de programação permite-nos explorar relações entre variáveis, identificar padrões, testar pressupostos sobre as observações e, ainda, comparar diferentes modelos explicativos, o que inclui a flexiblidade para combinar diferentes conjuntos de variáveis explicativas, possiblitando a identificação do modelo que melhor se adequa ao conjunto de dados.

Em suma, este trabalho não apenas aprofundou o nosso conhecimento estatístico, como também ressaltou a importância da utlização de ferramentas computacionais no âmbito da análise estatística.