Exercícios de Machine Learning

2023

Sumário

Introdução à Ciência de Dados Fundamentos e Aplicações - Morettin & Singer (exercícios do livro e alguns extras dados no curso Introdução ao Aprendizado de Máquina)

Questão 1 (Portfólio)

O formato preliminar dos estudos pode ser encontrado no GitHub: repositório. Materiais derivados destes estudos serão incorporados no site oficial em formato de artigos para um blog em josecarlosinfo, em específico as questões 3, 4 e 13.

Questão 2 (Revisão)

Capítulos 6 a 6.2, 6.3 e 6.5 - do livro de Morettin e Singer revisados.

Questão 3 (Morettin & Singer)

```r
library("dplyr") # Pacote de manipulação de dados

# Dados da tabela 6.13 do livro Introdução à Ciência de Dados (Morettin & Singer)
volume <- c(656, 692, 588, 799, 766, 800, 693, 602,
            737, 921, 923, 945, 816, 584, 642, 970)
peso <- c(630, 745, 690, 890, 825, 960, 835, 570,
          705, 955, 990, 725, 840, 640, 740, 945)
dados_lobo <- as_tibble(data.frame(volume, peso))
```

3.i

Considerando dados de volume (cm^3) e peso (g) do lobo direito do fígado de 16 pacientes submetidos a transplante inter-vivos, com o objetivo de estimar o peso por meio do volume, podemos usar um modelo de regressão linear simples. A suposta relação linear entre volume (variável independente) e o peso (variável resposta) pode ter a seguinte representação:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Onde, y é o peso do lobo direito do fígado, x é o volume, ϵ é o erro aleatório e os termos β são os parâmetros. A interpretação dos parâmetros nessa representação se dá por β_0 como sendo o valor do peso quando o volume é zero (caso mais extremo), e β_1 como sendo o "efeito" ou mudança esperada no peso para uma unidade de incremento do volume.

3.ii

Podemos notar que a **Figura 1** sugere uma relação linear positiva entre o volume e o peso, onde, quanto maior o volume, maior é o peso. É importante também observar no canto inferior direito um ponto que aparenta sair um pouco do comportamento linear geral dos dados.

```r
library("ggplot2") # Pacote para construção de gráficos

# Gráfico de dispersão entre peso e volume do lobo direito do fígado
ggplot(dados_lobo, aes(x = volume, y = peso)) +
  geom_point() +
  labs(x = expression("Volume (cm"^3*")"),
       y = expression("Peso (g)")) +
  ggtitle("Volume vs. Peso")
```

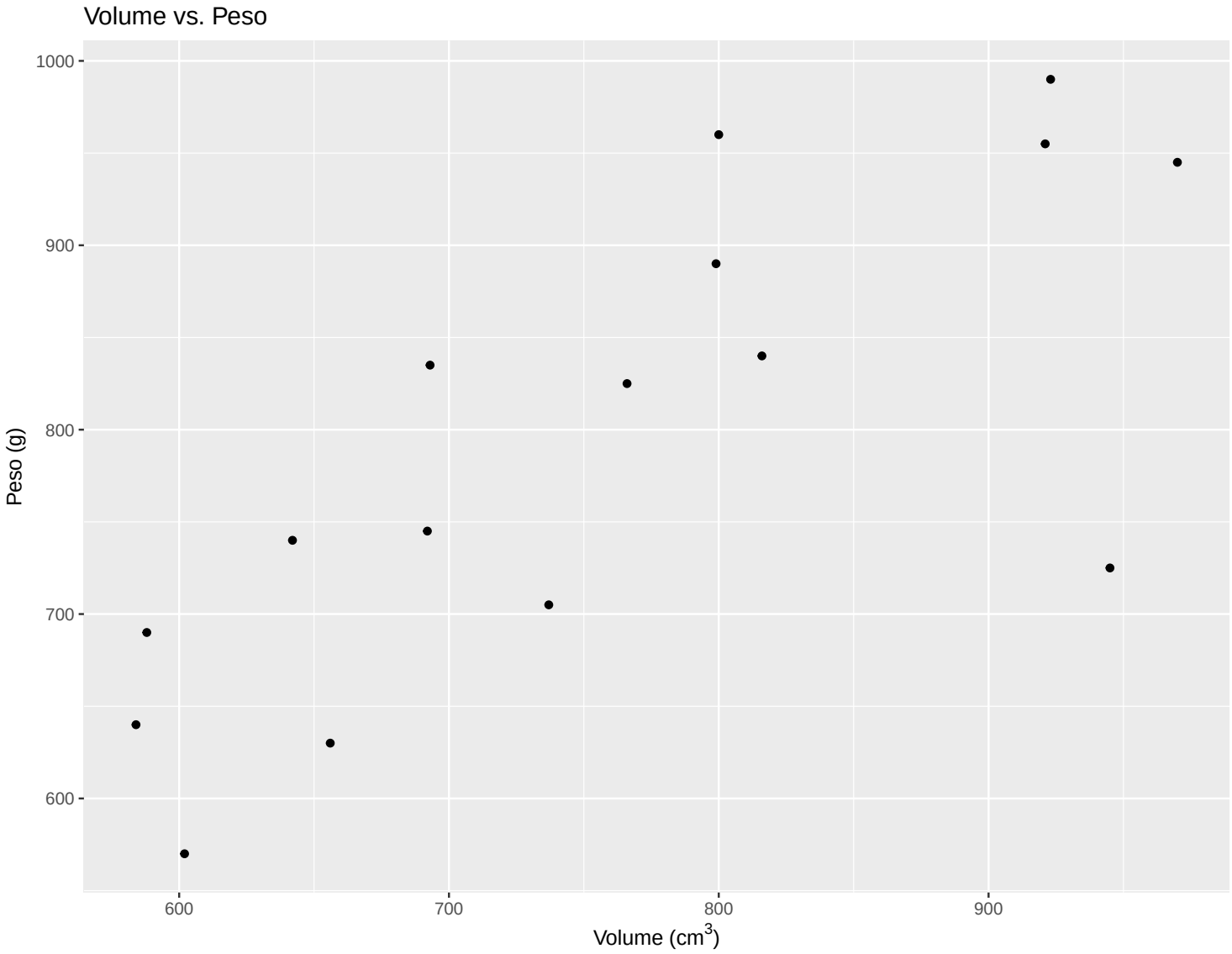


Figura 1: Gráfico de dispersão entre volume (cm^3) e peso (g) do lobo direito do fígado dos 16 pcientes.

3.iii

Ao ajustar o modelo linear simples, podemos ver no **Output 1** que a cada unidade de mudança no volume há um aumento de 0.76 (g) no peso do lobo direito do fígado dos pacientes. Além disso, olhando apenas para a métrica do coeficiente de determinação, o modelo explica 58% da variação do peso.

```
# Ajuste do modelo de regressão linear
modelo <- lm(peso ~ volume, data = dados_lobo)

# Resumo do modelo
summary(modelo)
```

```
##
## Call:
## lm(formula = peso ~ volume, data = dados_lobo)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -210.43  -32.54   14.76   44.97  135.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 213.2762   133.3334   1.600 0.132011
## volume        0.7642     0.1734   4.407 0.000597 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.91 on 14 degrees of freedom
## Multiple R-squared:  0.5811, Adjusted R-squared:  0.5512
## F-statistic: 19.42 on 1 and 14 DF,  p-value: 0.000597
```

Output 1: Resumo do modelo incluindo informações como coeficientes estimados, significância, etc.

Ao visualizar o ajuste pela **Figura 2**, o comportamento da reta em relação aos pontos me leva a suspeitar que possivelmente ela esteja sendo influenciada por algum ponto. Considerando que identificamos um em específico razoavelmente fora do padrão de comportamento dos dados, ele pode ser o culpado. Caso de fato seja, o correto é estudar esse ou os demais pontos que sejam influentes para determinar se faz sentido considerar a análise com ou sem eles incluídos.

```r
# Gráfico de dispersão dos dados originais com a linha de regressão
ggplot(dados_lobo, aes(x = volume, y = peso)) +
  geom_point() +  # Gráfico de dispersão dos dados
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Linha de regressão
  labs(x = "Volume Previsto (cm³)", y = "Peso Real (g)") +
  ggtitle("Ajuste do Modelo de Regressão Linear") +
  theme_minimal()
```

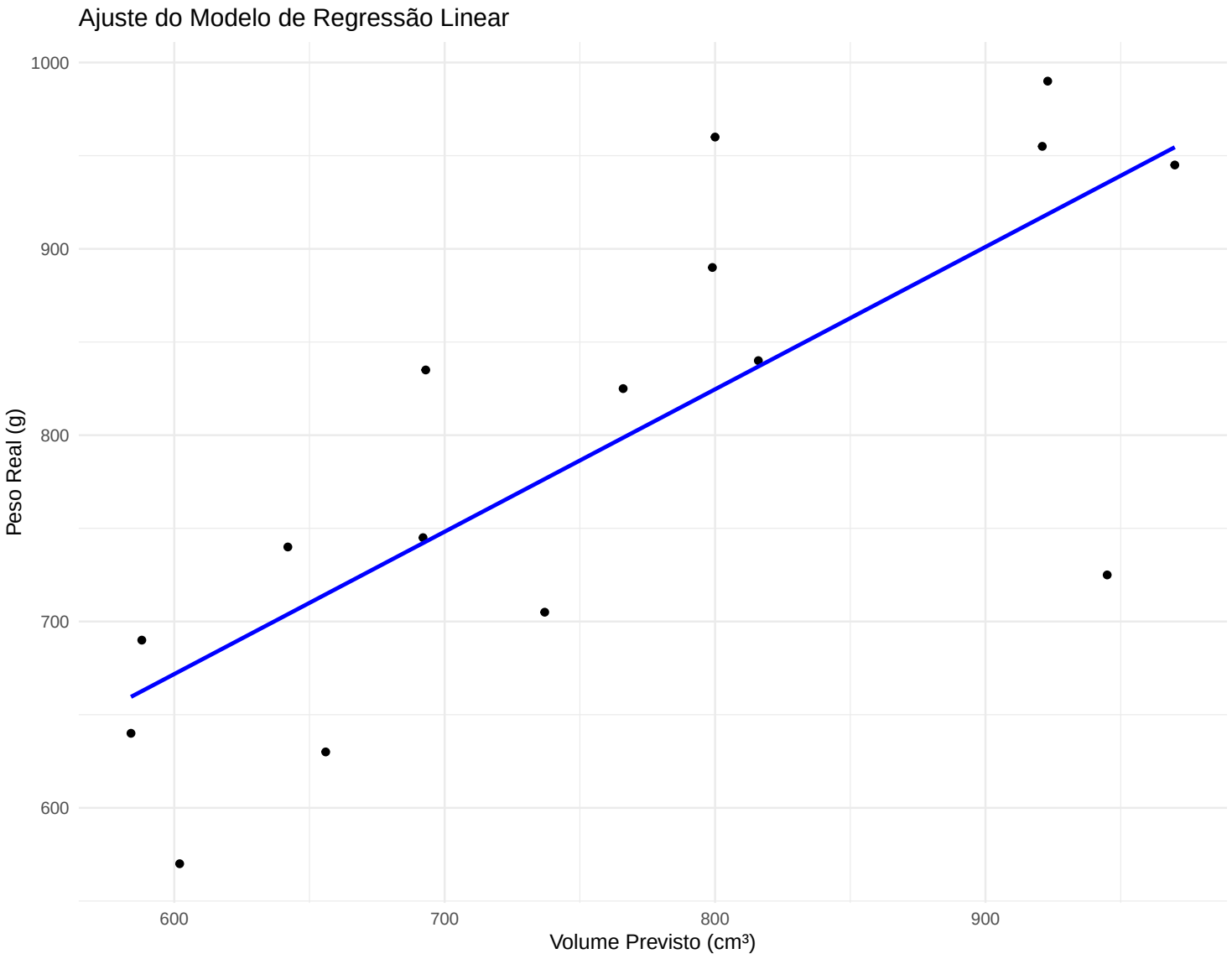


Figura 2: Gráfico do ajuste realizado.

Para verificar a suposição de normalidade dos resíduos, no **Output 2**, foram aplicados 6 diferentes testes de normalidade, onde, ao nível de 5% de significância, não há evidências para a rejeição da normalidade dos resíduos. Note que além dos testes de Shapiro, nenhum outro teve valor-p abaixo de 0.05, isso por que os testes de Shapiro são, em sua natureza, mais exigentes (leve em consideração que temos uma amostra pequena).

```
library("nortest") # Pacote para testes

# Função para testar normalidade
test_press <- function(k){
  rst <- rstudent(k)

require(nortest)

t1 <- ks.test(rst,"pnorm") #KS
t2 <- lillie.test(rst) # Lilliefors
t3 <- cvm.test(rst) # Cramér-von Mises
t4 <- shapiro.test(rst) # Shapiro-Wilk
t5 <- sf.test(rst) # Shapiro-Francia
```

```r
t6 <- ad.test(rst) # Anderson-Darling

# Tabela de resultados
testes <- c(t1$method,
            t2$method,
            t3$method,
            t4$method,
            t5$method,
            t6$method
            )
estt <- as.numeric(c(t1$statistic,
                     t2$statistic,
                     t3$statistic,
                     t4$statistic,
                     t5$statistic,
                     t6$statistic)
                   )
valorp <- c(t1$p.value, t2$p.value, t3$p.value, t4$p.value, t5$p.value,t6$p.value)
resultados <- cbind(estt, valorp)
rownames(resultados) <- testes
colnames(resultados) <- c("Estatística", "p")
print(resultados, digits = 4)

}

# Testando a normalidade dos resíduos
test_press(modelo)
```

```
##                                                Estatística       p
## Exact one-sample Kolmogorov-Smirnov test            0.1562 0.77513
## Lilliefors (Kolmogorov-Smirnov) normality test      0.1972 0.09694
## Cramer-von Mises normality test                     0.1102 0.07317
## Shapiro-Wilk normality test                         0.8716 0.02881
## Shapiro-Francia normality test                      0.8504 0.01560
## Anderson-Darling normality test                     0.6981 0.05482
```

Output 2: Testes de normalidade verificados e seus respectivos valores p.

Ao verificar a homocedasticidade, vemos no **Output 3** ao aplicar o teste de Breusch-Pagan, ao nível de 5% de significância, que não há evidências para a rejeição da hipótese de que os resíduos são distribuídos com igual variância (são homocedásticos). Além disso, ao aplicar o teste de Durbin-Watson, onde a hipótese nula é de que os resíduos não são correlacionados, ao nível de 5% de significância, não há evidências de correlação dos resíduos. Por fim, podemos ver que os resíduos padronizados estão todos contidos no intervalo -3 a 3, o que a princípio indica que não há presença de dados extremos (outliers).

```r
library("lmtest") # Pacote para testes de regressão
library("car") # Pacote para funções e testes de regressão

# Teste de Breusch-Pagan (Homocedasticidade)
bptest(modelo)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 1.5732, df = 1, p-value = 0.2097
```

```r
# Teste de Durbin-Watson (correlação dos resíduos)
durbinWatsonTest(modelo)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.0045856      1.923864   0.776
##  Alternative hypothesis: rho != 0
```

```r
# Verificando outliers
summary(rstandard(modelo))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.67306 -0.39452  0.17352 -0.01175  0.54902  1.59630
```

Output 3: Teste de Breusch-Pagan para homocedasticidade e de Durbin-Watson para verificar correlação dos resíduos. Além disso, a verificação de dados extremos utilizando resíduos padronizados.

3.iv

No **Output 4** foi computado os intervalos de confiança para os coeficientes estimados. Assim, temos uma estimativa de $\beta_0 = 213.27$ com intervalo de confiança de 95% sendo $[-72.70, 499.25]$, e de $\beta_1 = 0.76$ com intervalo de confiança de 95% sendo $[0.39, 1.14]$.

```r
# Construção dos intervalos de confiança para os parâmetros
intervalos_confianca <- confint(modelo)
print(intervalos_confianca)
```

```
##                   2.5 %     97.5 %
## (Intercept) -72.6955362 499.247847
## volume        0.3922543   1.136109
```

Output 4: Intervalos de confiança dos parâmetros do modelo.

3.v

A equação de regressão obtida é a seguinte:

$$\text{peso} = 213.27 + (0.76) \cdot \text{volume}$$

Considere a seguinte expressão para o intervalo de confiança do valor esperado:

$$\hat{y} \pm t_{\alpha/2,n-2}\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Levando em consideração os volumes 600, 700, 800, 900 e 1000 cm^3 sugeridos no enunciado. Teremos o seguinte:

Peso esperado ($\hat{y}$) para cada volume

$$\hat{y} = 213.27 + (0.76) \cdot 600 = 669.27$$
$$\hat{y} = 213.27 + (0.76) \cdot 700 = 745.27$$
$$\hat{y} = 213.27 + (0.76) \cdot 800 = 821.27$$

$$\hat{y} = 213.27 + (0.76) \cdot 900 = 897.27$$

$$\hat{y} = 213.27 + (0.76) \cdot 1000 = 973.27$$

Valor crítico da distribuição t ($t_{\alpha/2,n-2}$), o erro padrão ($\sqrt{MSE}$), a média amostral $\bar{x}$ e a soma $\sum (x_i - \bar{x})^2$

```
# Nível de confiança
nivel_confianca <- 0.95

# Valor crítico da distribuição t para o nível de confiança
(valor_critico <- qt((1 + nivel_confianca) / 2, df = nobs(modelo) - 2))
```

```
## [1] 2.144787
```

```
# Erro padrão
(erro_p <- summary(modelo)$sigma)
```

```
## [1] 87.90566
```

```
# Média amostral
(media_amostral <- mean(volume))
```

```
## [1] 758.375
```

```
# Soma (valor - média)^2
(soma_vm <- sum((volume - media_amostral)^2))
```

```
## [1] 256971.8
```

Output 5: Medidas necessárias para calcular os intervalos de confiança.

Calculando o intervalo considerando cada volume sugerido

```
# Calculando intervalos de confiança considerando cada volume sugerido

# Volume 600
seiscentos_superior <-
  669.27+valor_critico*erro_p*sqrt((1/16) + ((600 - media_amostral)^2)/soma_vm)
seiscentos_inferior <-
  669.27-valor_critico*erro_p*sqrt((1/16) + ((600 - media_amostral)^2)/soma_vm)

# Volume 700
setecentos_superior <-
  745.27+valor_critico*erro_p*sqrt((1/16) + ((700 - media_amostral)^2)/soma_vm)
setecentos_inferior <-
  745.27-valor_critico*erro_p*sqrt((1/16) + ((700 - media_amostral)^2)/soma_vm)

# Volume 800
oitocentos_superior <-
  821.27+valor_critico*erro_p*sqrt((1/16) + ((800 - media_amostral)^2)/soma_vm)
oitocentos_inferior <-
  821.27-valor_critico*erro_p*sqrt((1/16) + ((800 - media_amostral)^2)/soma_vm)

# Volume 900
```

```r
novecentos_superior <-
  897.27+valor_critico*erro_p*sqrt((1/16) + ((900 - media_amostral)^2)/soma_vm)
novecentos_inferior <-
  897.27-valor_critico*erro_p*sqrt((1/16) + ((900 - media_amostral)^2)/soma_vm)

# Volume 1000
mil_superior <-
  973.27+valor_critico*erro_p*sqrt((1/16) + ((1000 - media_amostral)^2)/soma_vm)
mil_inferior <-
  973.27-valor_critico*erro_p*sqrt((1/16) + ((1000 - media_amostral)^2)/soma_vm)

# Tabela
intervalo_superior <- c(round(seiscentos_superior, 2),
                        round(setecentos_superior, 2),
                        round(oitocentos_superior, 2),
                        round(novecentos_superior, 2),
                        round(mil_superior, 2))
intervalo_inferior <- c(round(seiscentos_inferior, 2),
                        round(setecentos_inferior, 2),
                        round(oitocentos_inferior, 2),
                        round(novecentos_inferior, 2),
                        round(mil_inferior, 2))
peso_estimado <- c(669.27, 745.27, 821.27, 897.27, 973.27)

tab_intervalos <- data.frame(peso = peso_estimado,
                             ICI = intervalo_inferior,
                             ICS = intervalo_superior)
print(tab_intervalos, row.names = FALSE)
```

```
##    peso    ICI     ICS
##  669.27 593.83  744.71
##  745.27 693.38  797.16
##  821.27 771.66  870.88
##  897.27 826.59  967.95
##  973.27 871.79 1074.75
```

Output 5: Tabela com os valores dos pesos esperados para cada volume e seus respectivos intervalos de confiança.

3.vi

O modelo sem intercepto terá a mesma representação apresentada no item 3.i, porém sem o termo β_0. Assim, a interpretação dos demais termos é a mesma.

Como não há nenhuma alteração nos dados, o item 3.ii também será o mesmo.

Ao ajustar o modelo linear simples sem intercepto, podemos ver no Output 1 que a cada unidade de mudança no volume há um aumento de 1.04 (g) no peso do lobo direito do fígado dos pacientes. Além disso, olhando apenas para a métrica do coeficiente de determinação, o modelo explica 98.7% da variação do peso (consideravelmente maior do que o R^2 do modelo com intercepto).

```r
# Ajuste do modelo de regressão linear sem intercepto
modelo_0 <- lm(peso ~ 0 + volume, data = dados_lobo)

# Resumo do modelo
summary(modelo_0)
```

```
##
## Call:
## lm(formula = peso ~ 0 + volume, data = dados_lobo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -255.69  -51.77   28.47   64.06  129.78
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## volume  1.03777    0.03003   34.56 1.03e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 92.36 on 15 degrees of freedom
## Multiple R-squared:  0.9876, Adjusted R-squared:  0.9868
## F-statistic:  1194 on 1 and 15 DF,  p-value: 1.026e-15
```

Output 6: Resumo do modelo sem intercepto incluindo informações como coeficientes estimados, significância, etc.

Ao visualizar o ajuste pela **Figura 3**, o comportamento da reta em relação aos pontos parece razoável, e a suspeita do ponto no canto inferior direto de influenciar a reta não existe mais.

```r
# Gráfico de dispersão dos dados originais com a linha de regressão
ggplot(dados_lobo, aes(x = volume, y = peso)) +
  geom_point() +  # Gráfico de dispersão dos dados
  geom_smooth(method = "lm",
              se = FALSE,
              color = "blue",
              formula = y ~ x - 1) +  # Linha de regressão (sem intercepto)
  labs(x = "Volume Previsto (cm³)", y = "Peso Real (g)") +
  ggtitle("Ajuste do Modelo de Regressão Linear (sem intercepto)") +
  theme_minimal()
```

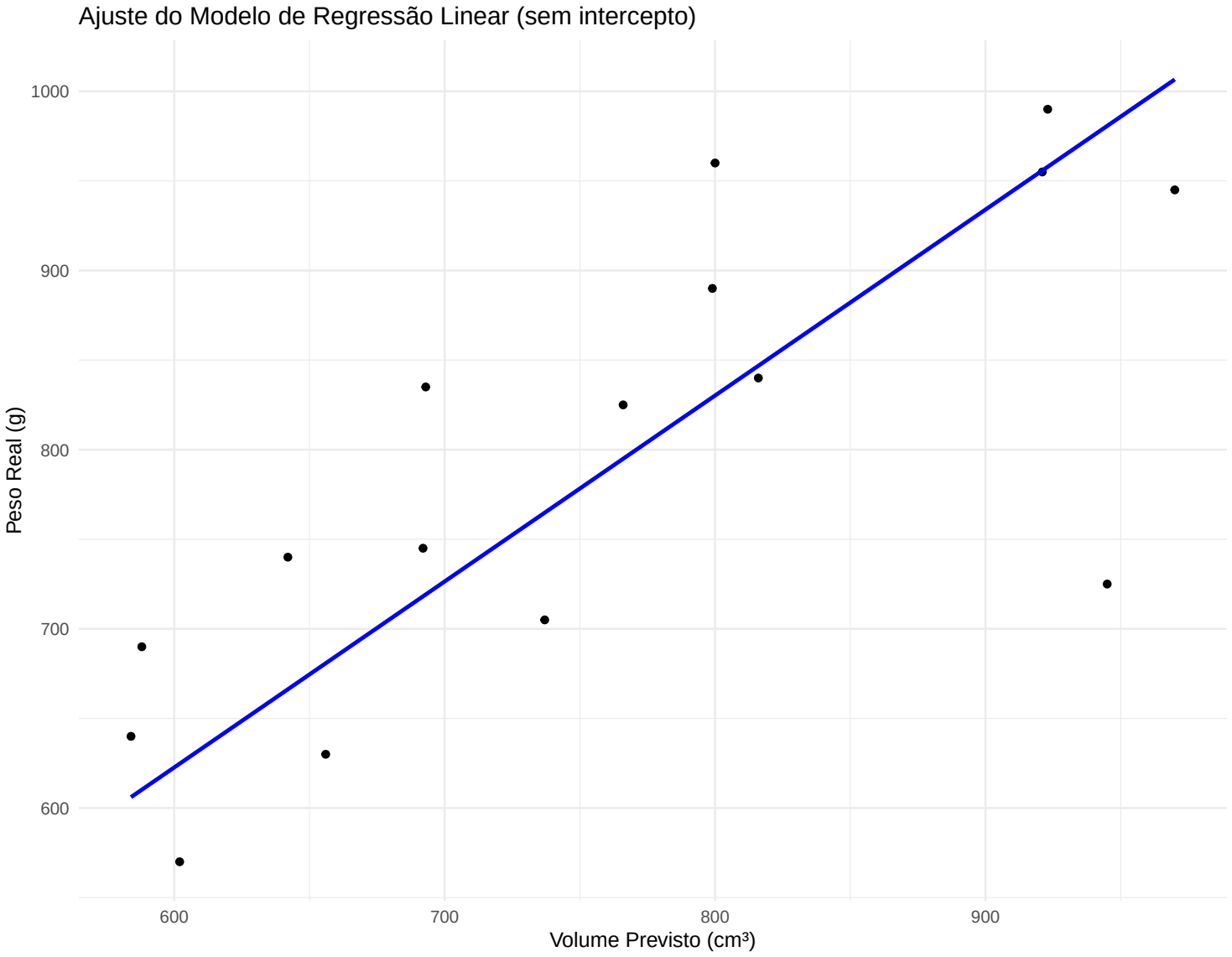


Figura 3: Gráfico do ajuste realizado sem intercepto.

Podemos obervar pelo **Output 7** que apenas o teste de Kolmogorov apontou para a normalidade dos resíduos. Além disso, O teste de Durbin-Watson apontou para resíduos não correlacionados, e ao verificar a presença de valores extremos através dos resíduos padronizados pude constatar que não há outliers. Por fim, não foi possível aplicar o teste de Breusch-Pagan para verificar homocedasticidade por ele requerir um modelo com intercepto, o que me levou a aplicar o teste de Goldfeld-Quandt, o qual apontou para resíduos homocedásticos. No geral, o modelo atende aos pressupostos, com a ressalva de que no quesito normalidade, apenas o teste de Kolmogorov o suporta.

```
# Testando a normalidade dos resíduos
test_press(modelo_0)
```

```
##                                             Estatística        p
## Exact one-sample Kolmogorov-Smirnov test         0.1823 0.5990808
## Lilliefors (Kolmogorov-Smirnov) normality test   0.2426 0.0125961
## Cramer-von Mises normality test                  0.1673 0.0120861
## Shapiro-Wilk normality test                      0.7708 0.0011454
## Shapiro-Francia normality test                   0.7446 0.0009648
## Anderson-Darling normality test                  1.1192 0.0044018
```

```r
# Teste de Goldfeld-Quandt (Homocedasticidade)
resultado_teste <- lmtest::gqtest(modelo_0, alternative = "two.sided")
print(resultado_teste)
```

```
##
##  Goldfeld-Quandt test
##
## data:  modelo_0
## GQ = 2.2044, df1 = 7, df2 = 7, p-value = 0.3188
## alternative hypothesis: variance changes from segment 1 to 2
```

```r
# Teste de Durbin-Watson (correlação dos resíduos)
durbinWatsonTest(modelo_0)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1141448       1.72187    0.55
##  Alternative hypothesis: rho != 0
```

```r
# Verificando outliers
summary(rstandard(modelo_0))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.90916 -0.57313  0.31733  0.06016  0.71559  1.45530
```

Output 7: Testes de diagnósticos do modelo sem intercepto.

No **Output 8** foi computado os intervalos de confiança para os coeficientes estimados. Assim, temos uma estimativa de $\beta_1 = 1.04$ com intervalo de confiança de 95% sendo $[0.97, 1.10]$.

```r
# Construção dos intervalos de confiança para os parâmetros
intervalos_confianca <- confint(modelo_0)
print(intervalos_confianca)
```

```
##            2.5 %   97.5 %
## volume 0.9737617 1.101777
```

Output 8: Intervalo de confiança do parâmetro do modelo sem intercepto.

A equação de regressão obtida é a seguinte:

$$\text{peso} = 0 + (1.04) \cdot \text{volume}$$

Irei considerar a mesma expressão para o intervalo de confiança do valor esperado utilizada no item 3.v. Levando em consideração os volumes 600, 700, 800, 900 e 1000 cm^3 sugeridos no enunciado. Teremos o seguinte:

Peso esperado ($\hat{y}$) para cada volume

$$\hat{y} = 0 + (1.04) \cdot 600 = 624$$
$$\hat{y} = 0 + (1.04) \cdot 700 = 728$$
$$\hat{y} = 0 + (1.04) \cdot 800 = 832$$
$$\hat{y} = 0 + (1.04) \cdot 900 = 936$$
$$\hat{y} = 0 + (1.04) \cdot 1000 = 1040$$

Valor crítico da distribuição t ($t_{\alpha/2,n-1}$), o erro padrão ($\sqrt{MSE}$), a média amostral $\bar{x}$ e a soma $\sum (x_i - \bar{x})^2$

```r
# Nível de confiança
nivel_confianca <- 0.95

# Valor crítico da distribuição t para o nível de confiança
(valor_critico0 <- qt((1 + nivel_confianca) / 2, df = nobs(modelo_0) - 1))
```

```
## [1] 2.13145
```

```r
# Erro padrão
(erro_p0 <- summary(modelo_0)$sigma)
```

```
## [1] 92.35988
```

```r
# Média amostral
(media_amostral <- mean(volume))
```

```
## [1] 758.375
```

```r
# Soma (valor - média)^2
(soma_vm <- sum((volume - media_amostral)^2))
```

```
## [1] 256971.8
```

Output 9: Medidas necessárias para calcular os intervalos de confiança considerando o modelo sem intercepto.

Calculando o intervalo considerando cada volume sugerido

```r
# Calculando intervalos de confiança considerando cada volume sugerido

# Volume 600
seiscentos_superior0 <-
  624+valor_critico0*erro_p0*sqrt((1/16) + ((600 - media_amostral)^2)/soma_vm)
seiscentos_inferior0 <-
  624-valor_critico0*erro_p0*sqrt((1/16) + ((600 - media_amostral)^2)/soma_vm)

# Volume 700
setecentos_superior0 <-
  728+valor_critico0*erro_p0*sqrt((1/16) + ((700 - media_amostral)^2)/soma_vm)
setecentos_inferior0 <-
  728-valor_critico0*erro_p0*sqrt((1/16) + ((700 - media_amostral)^2)/soma_vm)

# Volume 800
oitocentos_superior0 <-
  832+valor_critico0*erro_p0*sqrt((1/16) + ((800 - media_amostral)^2)/soma_vm)
oitocentos_inferior0 <-
  832-valor_critico0*erro_p0*sqrt((1/16) + ((800 - media_amostral)^2)/soma_vm)

# Volume 900
novecentos_superior0 <-
  936+valor_critico0*erro_p0*sqrt((1/16) + ((900 - media_amostral)^2)/soma_vm)
novecentos_inferior0 <-
  936-valor_critico0*erro_p0*sqrt((1/16) + ((900 - media_amostral)^2)/soma_vm)
```

```r
# Volume 1000
mil_superior0 <-
  1040+valor_critico0*erro_p0*sqrt((1/16) + ((1000 - media_amostral)^2)/soma_vm)
mil_inferior0 <-
  1040-valor_critico0*erro_p0*sqrt((1/16) + ((1000 - media_amostral)^2)/soma_vm)

# Tabela
intervalo_superior0 <- c(round(seiscentos_superior0, 2),
                         round(setecentos_superior0, 2),
                         round(oitocentos_superior0, 2),
                         round(novecentos_superior0, 2),
                         round(mil_superior0, 2))
intervalo_inferior0 <- c(round(seiscentos_inferior0, 2),
                         round(setecentos_inferior0, 2),
                         round(oitocentos_inferior0, 2),
                         round(novecentos_inferior0, 2),
                         round(mil_inferior0, 2))
peso_estimado0 <- c(624, 728, 832, 936, 1040)

tab_intervalos0 <- data.frame(peso = peso_estimado0,
                              ICI = intervalo_inferior0,
                              ICS = intervalo_superior0)
print(tab_intervalos0, row.names = FALSE)
```

```
##  peso    ICI     ICS
##   624 545.23  702.77
##   728 673.81  782.19
##   832 780.20  883.80
##   936 862.20 1009.80
##  1040 934.04 1145.96
```

Output 10: Tabela com os valores dos pesos esperados para cada volume e seus respectivos intervalos de confiança considerando o modelo sem intercepto.

No geral o modelo sem intecepto está bem ajustado, com algumas ressalvas a serem feitas a respeito da normalidade dos resíduos. O problema é que esse modelo automaticamente assume que se o volume do lobo direito do fígado for zero, então o peso também será zero, o que é uma suposição muito forte a ser feita, e deve se ter muito cuidado. No geral o intercepto é sempre incluído na análise, embora provavelmente possa existir casos específicos nos quais não necessitem, e como não tenho certeza se de fato faz sentido assumir o intercepto zero nesse cenário (consultar com profissionais da área ou pesquisador responsável), o modelo com intercepto acaba sendo mais conveniente.

Questão 4 (Morettin & Singer)

A planilha chamada "dados_pico," que aparece no **Output 11**, foi criada por mim no Excel, selecionando apenas as variáveis de interesse na questão. Ela contém informações de perfil e medidas obtidas no momento de pico do exercício.

```r
library("readxl") # Pacote para importar dados
library("knitr") # Tabelas com melhor formatação para LateX

# Lista planilhas do arquivo
excel_sheets("esforco.xls")
```

```
## [1] "descricao"  "dados_pico" "dados"
```

```r
# Carrega os dados
dados_0 <- read_excel("esforco.xls", sheet = "dados_pico")

# Primeiras 5 linhas da base de dados
linhas_5 <- head(dados_0, 5)
kable(linhas_5, format = "latex")
```

id	aw	au	k	av	ax	h	d	f
1	14.1	71	2	118	1.26	54	M	38
2	16.3	91	1	113	1.09	80	M	49
3	9.9	37	2	148	1.10	56	F	65
4	17.7	127	2	144	1.34	78	M	52
5	10.8	43	4	107	1.06	59	F	52

Output 11: Importação dos dados de interesse e visualização das primeiras 5 linhas.

O significado de cada variável é apresentado abaixo:

id: identificador de cada indivíduo.

aw: consumo de oxigênio em ml/(kg.min) no pico do exercício (Desfecho).

au: carga na esteira ergométrica.

k: classe funcional pelo critério NYHA (1 a 4). É uma classificação usada para predizer o prognóstico e a sobrevida de pacientes com insuficiência cardíaca.

av: frequência cardíaca (bpm).

ax: razão de troca respiratória.

h: peso (kg).

d: sexo (F: feminino, M: masculino).

f: idade (anos).

Tratamento dos dados

No **Output 12**, podemos ter uma noção da estrutura da base de dados. No R, é interessante lidar com variáveis categóricas como sendo fatores, podendo acrescentar ou não ordem às classes da variável. Assim, dentre as variáveis da base de dados, as variáveis **k** (cada nível da classificação implica uma progressão na gravidade dos sintomas da insuficiência cardíaca) e **d** (sem ordem aparente) devem ser transformadas em fatores. Além disso, não foram identificados dados faltantes (pelo menos não como dados faltantes do R).

```r
# Estrutura da base de dados
dplyr::glimpse(dados_0)
```

```
## Rows: 127
## Columns: 9
## $ id <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, ~
## $ aw <dbl> 14.1, 16.3, 9.9, 17.7, 10.8, 14.0, 9.5, 13.9, 11.8, 18.1, 16.8, 15.~
## $ au <chr> "71", "91", "37", "127", "43", "60", "32", "63", "71", "112", "95",~
## $ k  <dbl> 2, 1, 2, 2, 4, 1, 3, 2, 3, 1, 3, 3, 1, 4, 2, 2, 2, 2, 2, 2, 3, 3~
## $ av <dbl> 118, 113, 148, 144, 107, 135, 117, 147, 175, 148, 184, 141, 164, 10~
## $ ax <dbl> 1.26, 1.09, 1.10, 1.34, 1.06, 1.12, 1.27, 1.28, 1.16, 1.23, 1.35, 1~
## $ h  <dbl> 54, 80, 56, 78, 59, 62, 42, 55, 77, 81, 69, 51, 70, 52, 98, 67, 58,~
## $ d  <chr> "M", "M", "F", "M", "F", "F", "F", "F", "F", "M", "M", "M", "M", "M~
## $ f  <dbl> 38, 49, 65, 52, 52, 58, 24, 39, 48, 50, 45, 59, 34, 56, 62, 45, 62,~
```

```r
# Número de dados faltantes em cada coluna
colSums(is.na(dados_0))
```

```
## id aw au  k av ax  h  d  f
##  0  0  0  0  0  0  0  0  0
```

Output 12: Classe de cada coluna no R e quantidade de dados faltantes.

Ao verificar as classes da variável **k**, vemos no **Output 13** que ela está apresentando as classes corretas de 1 a 4, mas também possui uma classe 0. Observe que as 5 primeiras linhas da base, onde apenas as linhas com classe 0 foram selecionadas, possuem exatamente os mesmos valores de identificação da variável **id** apresentados no **Output 11**, mas neste último os valores de **k** não eram zero. Ao investigar a base de dados, foi possível notar que os identificadores são únicos apenas até o 87; as 40 linhas a partir dele possuem os identificadores em ordem de 1 a 40, ou seja, os identificadores dos 40 primeiros indivíduos estão duplicados. Porém, a duplicidade se apresenta apenas nos identificadores, enquanto que as medidas das demais variáveis diferem. Assim, é difícil saber se essas pessoas foram observadas duas vezes, ou se houve algum engano ao preencher a informação de identificação, sendo este último o mais provável, visto que as características físicas e medidas das variáveis diferem, mas fato que não justifica valores zero no critério NYHA.

```r
# Valores presentes na coluna "k"
unique(dados_0$k)
```

```
## [1] 2 1 4 3 0
```

```r
# Base de dados onde as linhas são k = 0 (primeiras 5 linhas)
linhak_5 <- head(dados_0[dados_0$k == 0,], 5)
kable(linhak_5, format = "latex")
```

id	aw	au	k	av	ax	h	d	f
1	18.3	108	0	168	1.30	86	M	69
2	14.7	141	0	142	1.58	109	M	69
3	18.7	105	0	151	1.20	75	M	71
4	28.6	135	0	135	1.27	60	M	63
5	24.0	148	0	134	1.18	82	M	61

Output 13: Identificação das linhas com classificação NYHA igual a zero.

Considerando que o critério NYHA possui apenas 4 níveis (de 1 a 4), decidi trabalhar apenas com os dados dos quais existe certeza da informação apresentada, ou seja, apenas casos em que o critério NYHA (variável **k**) é válido. O tratamento é realizado no **Output 14**.

```r
# Base de dados sem as linhas com classificação k = 0
dados_1 <- dados_0[dados_0$k != 0,]

# Transforma k em fator ordinal
ordem_valores <- c(1, 2, 3, 4)
ordem_categorias <- c("classe1", "classe2", "classe3", "classe4")
dados_1$k <- factor(dados_1$k,
                    levels = ordem_valores,
                    labels = ordem_categorias)

# Transforma d em fator sem ordem
dados_1$d <- factor(dados_1$d, ordered = FALSE)
```

Output 14: Base de dados com apenas as classificações válidas de NYHA e transformando as variáveis **k** e **d** em fatores.

Note no **Output 15** que a variável **au** está sendo interpretada como texto, o que não faz sentido, visto que essa variável se refere à carga na esteira ergométrica. Ao verificar os valores únicos dessa coluna, foi possível identificar um caractere `"."`, o qual está

fazendo com que toda a coluna seja do tipo texto. Nesse caso, como são apenas duas linhas com esse problema, decidi imputar a média dos valores, o que é razoável, considerando que a média e mediana são próximas.

```
# Valores presentes na coluna "au"
unique(dados_1$au)
```

```
##  [1] "71"  "91"  "37"  "127" "43"  "60"  "32"  "63"  "112" "95"  "115" "33"
## [13] "100" "105" "84"  "94"  "75"  "."   "68"  "15"  "56"  "101" "97"  "93"
## [25] "145" "23"  "104" "62"  "142" "85"  "218" "79"  "76"  "125" "66"  "110"
## [37] "45"  "77"  "153" "113" "82"  "69"  "98"  "67"  "52"  "114" "40"  "65"
## [49] "55"  "48"  "58"  "135" "11"  "86"  "154"
```

```
aux_au <- dados_1[dados_1$au != ".","au"] # Vetor de dados númericos
summary(as.numeric(aux_au$au)) # Sumário
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   62.00   86.00   86.59  105.00  218.00
```

```
mean_au <- mean(as.numeric(aux_au$au)) # Média

dados_1[dados_1$au == ".","au"] <- as.character(mean_au) # Troca "." pela média
dados_1$au <- as.numeric(dados_1$au) # Classe correta da variável
```

Output 15: Tratamento da variável **au**.

Por fim, pode-se visualizar a estrutura final da base de dados no **Output 16**, onde todas as variáveis estão com suas classes corretas.

```
# Estrutura da base de dados
dplyr::glimpse(dados_1)
```

```
## Rows: 87
## Columns: 9
## $ id <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, ~
## $ aw <dbl> 14.1, 16.3, 9.9, 17.7, 10.8, 14.0, 9.5, 13.9, 11.8, 18.1, 16.8, 15.~
## $ au <dbl> 71.00000, 91.00000, 37.00000, 127.00000, 43.00000, 60.00000, 32.000~
## $ k  <fct> classe2, classe1, classe2, classe2, classe4, classe1, classe3, clas~
## $ av <dbl> 118, 113, 148, 144, 107, 135, 117, 147, 175, 148, 184, 141, 164, 10~
## $ ax <dbl> 1.26, 1.09, 1.10, 1.34, 1.06, 1.12, 1.27, 1.28, 1.16, 1.23, 1.35, 1~
## $ h  <dbl> 54, 80, 56, 78, 59, 62, 42, 55, 77, 81, 69, 51, 70, 52, 98, 67, 58,~
## $ d  <fct> M, M, F, M, F, F, F, F, F, M, M, M, M, M, M, M, M, M, M, M, M, F~
## $ f  <dbl> 38, 49, 65, 52, 52, 58, 24, 39, 48, 50, 45, 59, 34, 56, 62, 45, 62,~
```

Output 16: Estrutura da base de dados após o tratamento.

4.a

Considerando o desfecho de consumo de oxigênio em ml/(kg.min) no pico do exercício (variável **aw**), podemos observar na **Figura 4** que apenas a carga na esteira ergométrica (variável **au**) e a frequência cardíaca (variável **av**) possuem uma correlação moderada a forte com o desfecho, apresentando uma correlação de Pearson de 0.758 e 0.490, respectivamente.

Desconsiderando o desfecho, observa-se que a variável **au** possui uma correlação não desprezível com as variáveis **av** e **h**, sendo de 0.360 e 0.424 respectivamente. As demais correlações são todas fracas ou ligeiramente baixas para serem consideradas relevantes.

```r
library("GGally") # Extenção para o ggplot2

# Crie a matriz de gráficos de dispersão
scatter_matrix <- ggpairs(
    dados_1[, c("aw", "au", "av", "ax", "h", "f")],
    lower = list(continuous = "points", combo = "box")
)

# Imprima a matriz de gráficos de dispersão ajustada
print(scatter_matrix)
```

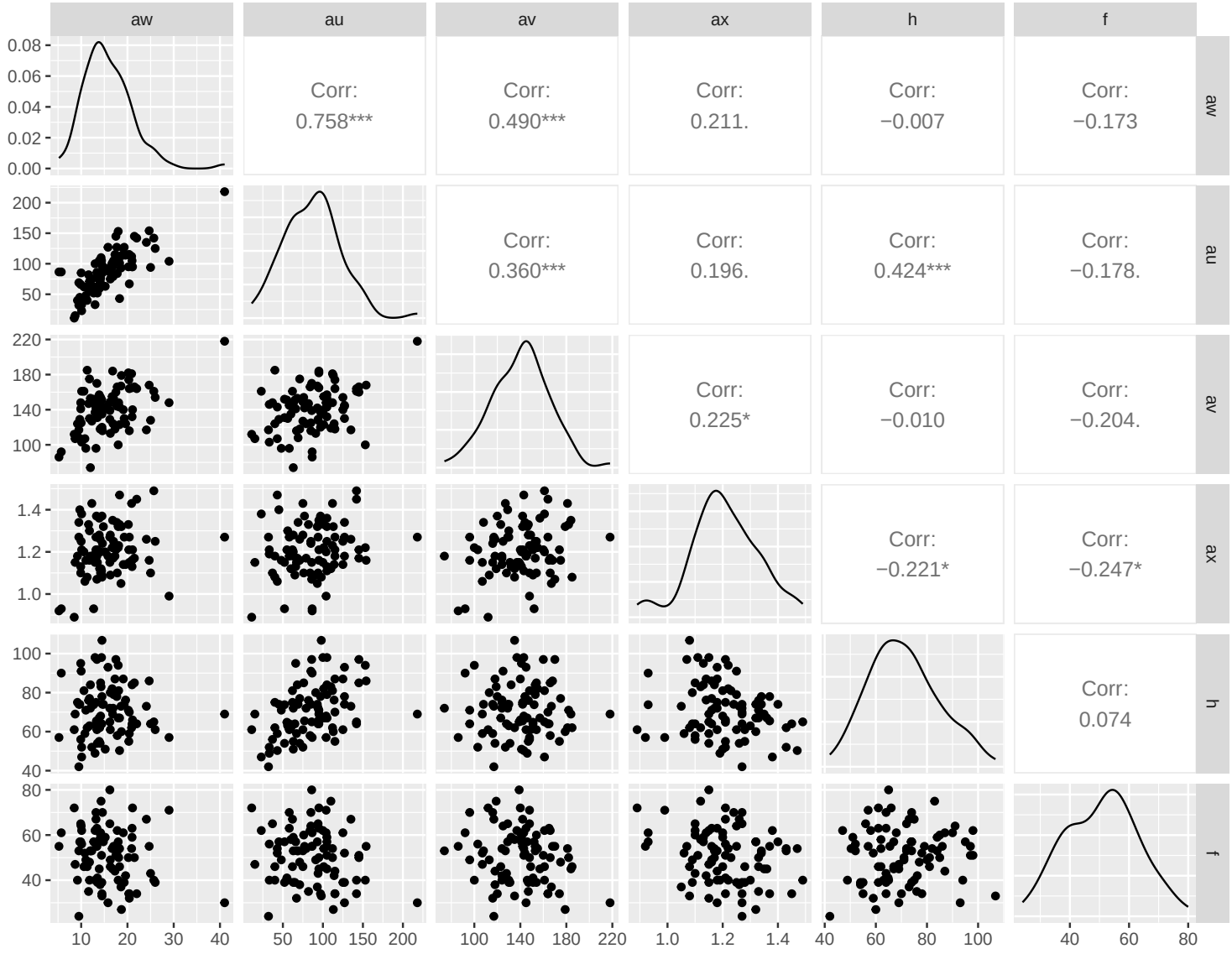


Figura 4: Matriz de gráficos de dispersão das variáveis numéricas. Na diagonal, estão as curvas de densidade das distribuições de cada variável numérica. Na parte inferior da matriz estão os gráficos de dispersão 2 x 2, e na parte superior, a correlação de Pearson.

Agora, considerando os dados agrupados pelos níveis do critério de NYHA, a **Figura 5** apresenta a matriz dos gráficos de dispersão com informações agrupadas. Considerando que **au** e **av** foram as variáveis com maior correlação de Pearson com o desfecho (relação linear), podemos observar que, ao analisar os níveis do critério de NYHA, as classes de 1 a 3 apresentam uma

correlação mais relevante com **au**, e uma correlação baixa, porém não desprezível, na classe 4. Enquanto isso, para **av**, ocorre uma maior correlação nas classes 1 e 3, e uma relação não desprezível na classe 2.

Entre as variáveis, sem considerar o desfecho, o destaque está nas correlações com a variável **au**, sendo a classe 1 do **av** (correlação de 0.683), a classe 2 do **ax** (correlação de 0.566), as classes 2 e 4 do **h** (correlação de 0.508 e 0.637, respectivamente) e as classes 1 e 2 do **f** (correlação de -0.535 e -0.459, respectivamente). Além disso, uma correlação possivelmente relevante pode ser observada entre **av** e **f** ao analisar a classe 1, apresentando uma correlação de -0.563.

```r
# Crie uma cópia dos dados sem a coluna "k"
dados_sem_k <- dados_1[, c("aw", "au", "av", "ax", "h", "f")]

# Crie a matriz de gráficos de dispersão
scatter_matrix <- ggpairs(
    dados_sem_k,
    ggplot2::aes(colour = dados_1$k),
    lower = list(continuous = "points", combo = "box")
)

# Imprima a matriz de gráficos de dispersão ajustada
print(scatter_matrix)
```

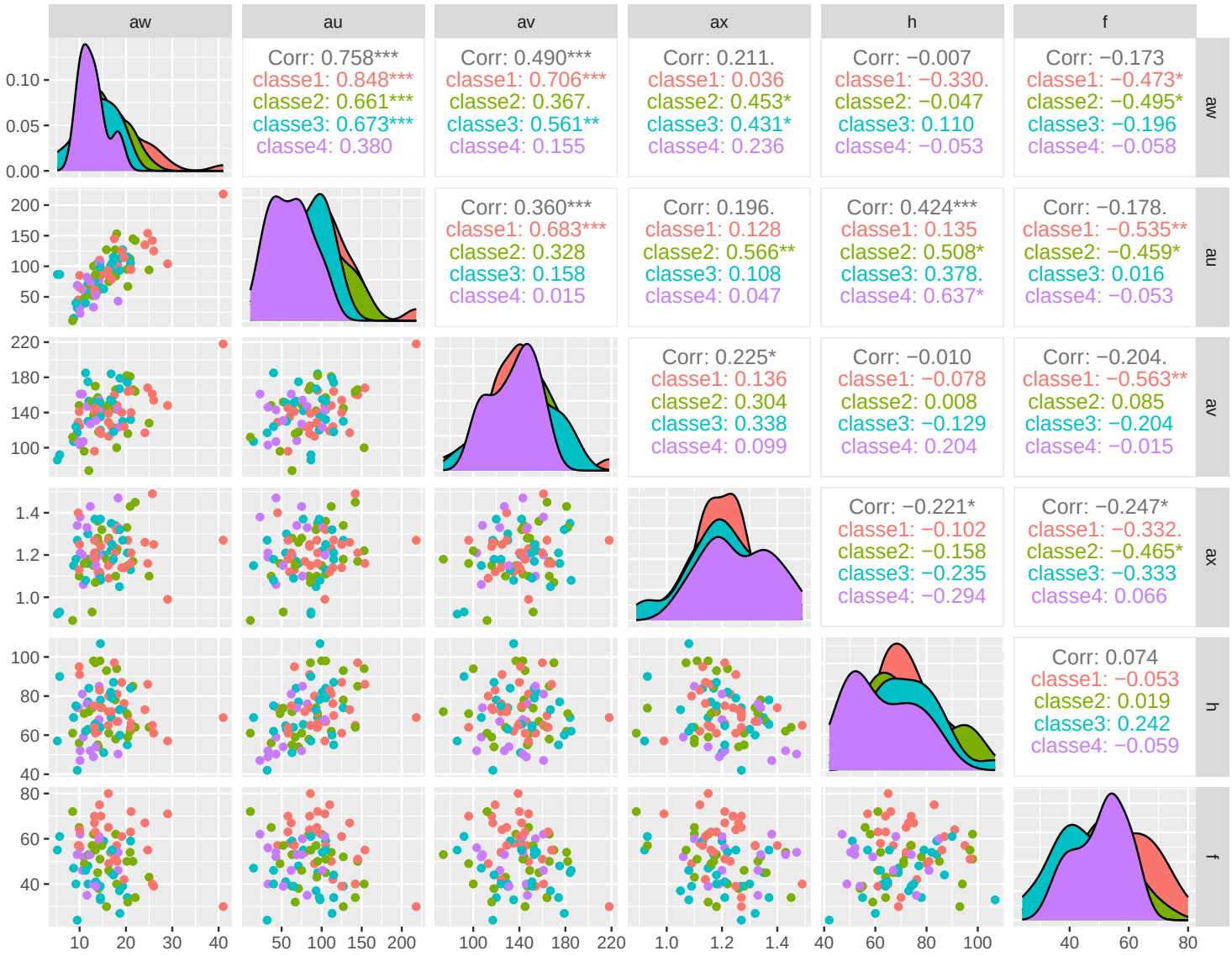


Figura 5: Matriz de gráficos de dispersão das variáveis numéricas, agora considerando os níveis do critério NYHA (variável **k**). Na diagonal, estão as curvas de densidade das distribuições de cada variável numérica agrupada. Na parte inferior da matriz estão os gráficos de dispersão 2 x 2 agrupados, e na parte superior, a correlação de Pearson.

Por fim, a **Figura 6** considera os dados agrupados pela variável **sexo**. A primeira coisa que é possível notar graças às cores nos gráficos de dispersão é que há consideravelmente mais indivíduos do sexo masculino do que do feminino na amostra. Outro ponto é que a distribuição do consumo de oxigênio (**aw**) e carga na esteira ergométrica (**au**) é bem diferente entre os grupos, o que não parece ocorrer nas demais variáveis.

Olhando para a correlação das variáveis **au** e **av** com o desfecho, o sexo masculino apresenta a maior correlação, sendo 0.751 e 0.515, respectivamente (observação: 0.414 para o sexo feminino no **av** não é desprezível).

Desconsiderando o desfecho, o destaque se encontra na classe masculina para uma correlação não desprezível entre **au** e **av**, sendo de 0.391. Além disso, em ambas as classes na correlação entre **au** e **h**, sendo de 0.814 (masculino) e 0.400 (feminino).

```
# Crie uma cópia dos dados sem a coluna "d"
dados_sem_d <- dados_1[, c("aw", "au", "av", "ax", "h", "f")]

# Crie a matriz de gráficos de dispersão
scatter_matrix <- ggpairs(
    dados_sem_d,
```

```
    ggplot2::aes(colour = dados_1$d),
    lower = list(continuous = "points", combo = "box")
)

# Imprima a matriz de gráficos de dispersão ajustada
print(scatter_matrix)
```

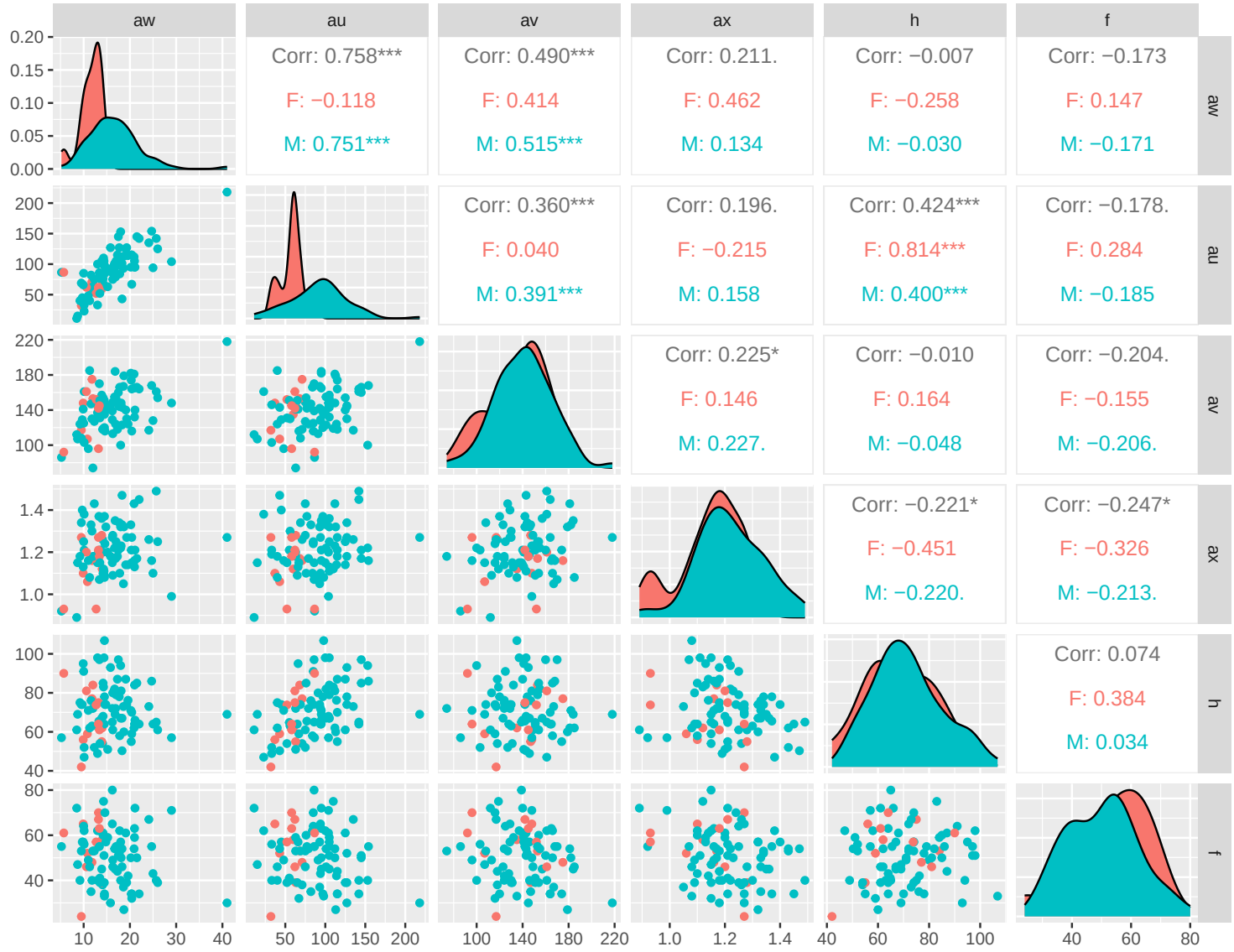


Figura 6: Matriz de gráficos de dispersão das variáveis numéricas, agora considerando a variável **sexo**. Na diagonal, estão as curvas de densidade das distribuições de cada variável numérica agrupada. Na parte inferior da matriz estão os gráficos de dispersão 2 x 2 agrupados, e na parte superior, a correlação de Pearson.

4.b

Com o objetivo de ajustar um modelo linear tendo o consumo de oxigênio no pico do exercício (**aw**) dos pacientes com insuficiência cardíaca como desfecho, podemos usar um modelo de regressão linear múltiplo. A suposta relação linear entre as variáveis independentes e **aw** (variável resposta) pode ter a seguinte representação:

$$\text{aw} = \beta_0 + \beta_1 \text{au} + \beta_2 \text{av} + \beta_3 \text{ax} + \beta_4 \text{h} + \beta_5 \text{f} + \beta_6 \text{dM} + \beta_7 \text{k2} + \beta_8 \text{k3} + \beta_9 \text{k4} + \epsilon$$

Onde **aw** é o desfecho, e **au**, **av**, **ax**, **h**, **f**, **d**, e **k** são as variáveis independentes. Os termos β representam os parâmetros e ϵ é o erro aleatório. A interpretação dos parâmetros nessa representação se dá da seguinte forma:

β_0 é o consumo esperado de oxigênio no pico do exercício quando todas as variáveis independentes são zero (caso mais extremo). β_1 é o "efeito" ou mudança esperada no consumo de oxigênio no pico do exercício para um incremento de uma unidade na carga na esteira ergométrica. Similarmente, β_2 é a mudança esperada no consumo de oxigênio no pico do exercício para um incremento de uma unidade na frequência cardíaca, e assim por diante para as demais covariáveis numéricas.

No caso da covariável **d**, que se refere ao sexo, uma das duas categorias é considerada como referência. Nessa estrutura, a categoria feminina é a referência. Portanto, β_6 representa a mudança esperada no consumo de oxigênio no pico do exercício vindo da categoria masculina em comparação com a categoria feminina, mantendo as outras covariáveis constantes. A mesma ideia se aplica à variável **k**, onde uma categoria (no caso, a classe 1) é usada como referência. Dessa forma, β_7 é a mudança esperada no consumo de oxigênio vindo da classe 2 em comparação com a classe 1, e assim sucessivamente para β_8 e β_9, considerando as classes 3 e 4, respectivamente.

4.c

No **Output 17** é apresentado o modelo de regressão linear múltiplo considerando todas as covariáveis. A informação do erro padrão e da significância dos parâmetros também é apresentada, onde as variáveis menos significativas (considerando nível de 5% de significância) são a razão de troca respiratória (**ax**), idade (**f**) e sexo (**d**).

```r
# Ajuste um modelo de regressão linear múltipla
modelo_0 <- lm(aw ~ au + av + ax + h + f + d + k, data = dados_1)

# Resumo do modelo
summary(modelo_0)
```

```
##
## Call:
## lm(formula = aw ~ au + av + ax + h + f + d + k, data = dados_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3968  -1.2644  -0.2467   1.0957   6.8689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.56970    4.76682   3.896 0.000207 ***
## au           0.11250    0.01187   9.481 1.42e-14 ***
## av           0.04737    0.01283   3.692 0.000413 ***
## ax          -5.16636    2.73814  -1.887 0.062955 .
## h           -0.15289    0.02570  -5.950 7.45e-08 ***
## f           -0.03568    0.02950  -1.210 0.230160
## dM           1.70194    0.87969   1.935 0.056700 .
## kclasse2    -1.77695    0.82744  -2.148 0.034898 *
## kclasse3    -3.05963    0.91616  -3.340 0.001295 **
## kclasse4    -2.58969    1.04744  -2.472 0.015627 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.68 on 77 degrees of freedom
## Multiple R-squared:  0.7819, Adjusted R-squared:  0.7564
## F-statistic: 30.67 on 9 and 77 DF,  p-value: < 2.2e-16
```

Output 17: Modelo de regressão linear múltiplo considerando todas as covariáveis da base de dados.

4.d

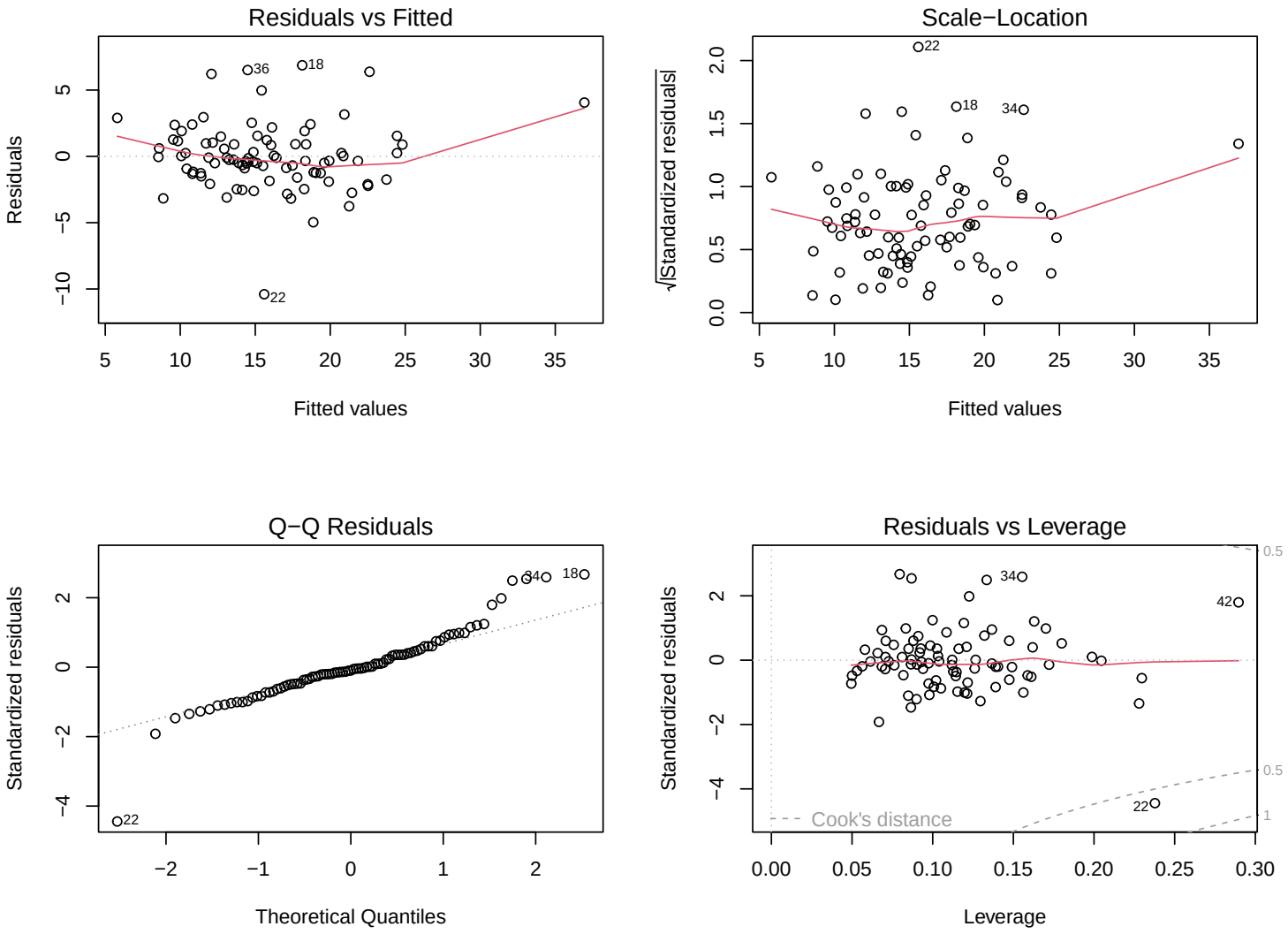
O **Output 18** apresenta quatro gráficos de diagnóstico básicos para a regressão linear. Os gráficos **Residuals vs Fitted** e **Scale-Location** não apresentam nenhum padrão aparente nos pontos, o que indica linearidade nos parâmetros estimados e homocedasticidade dos resíduos.

Além disso, o **Q-Q Residuals** segue bem a reta para a maioria dos pontos. No entanto, na cauda superior, os pontos se afastam consideravelmente, especialmente as observações 34 e 18. Na cauda inferior, a maior preocupação está na observação 22.

Por fim, **Residuals vs Leverage** aponta a observação 22 a uma distância relativamente alta de Cook (superior a 0.5), embora não seja extrema (é inferior a 1).

```
# Layout
layout(matrix(c(1, 2, 3, 4), 2, 2))

# Gráficos de diagnóstico
plot(modelo_0)
```



Output 18: Gráficos de diagnósticos: **Residuals vs Fitted** para verificar linearidade; **Scale-Location** para verificar homocedasticidade; **Q-Q Residuals** para verificar normalidade e **Residuals vs Leverage** para verificar pontos influentes.

4.e

Como vimos anteriormente no **Output 17**, considerando o nível de 5% de significância, apenas as variáveis **ax** (valor-p de 0.063), **f** (valor-p de 0.23) e **d** (valor-p de 0.057) não foram significativas. Portanto, serão removidas do modelo.

4.f

Ajuste

Podemos notar **Output 19** que todas as variáveis são significativas com exceção da classe 2 da variável **k**.

```
# Ajuste um modelo de regressão linear múltipla
modelo_1 <- lm(aw ~ au + av + h + k, data = dados_1)

# Resumo do modelo
summary(modelo_1)
```

```
##
## Call:
## lm(formula = aw ~ au + av + h + k, data = dados_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2712 -1.5774 -0.0827  1.1675  7.8047
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.12995    2.49816   4.455 2.69e-05 ***
## au           0.11861    0.01066  11.127  < 2e-16 ***
## av           0.04359    0.01298   3.358  0.00120 **
## h           -0.14641    0.02499  -5.860 9.89e-08 ***
## kclasse2    -1.29370    0.79000  -1.638  0.10543
## kclasse3    -2.33271    0.80295  -2.905  0.00474 **
## kclasse4    -2.29680    1.00735  -2.280  0.02527 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.747 on 80 degrees of freedom
## Multiple R-squared:  0.7618, Adjusted R-squared:  0.7439
## F-statistic: 42.64 on 6 and 80 DF,  p-value: < 2.2e-16
```

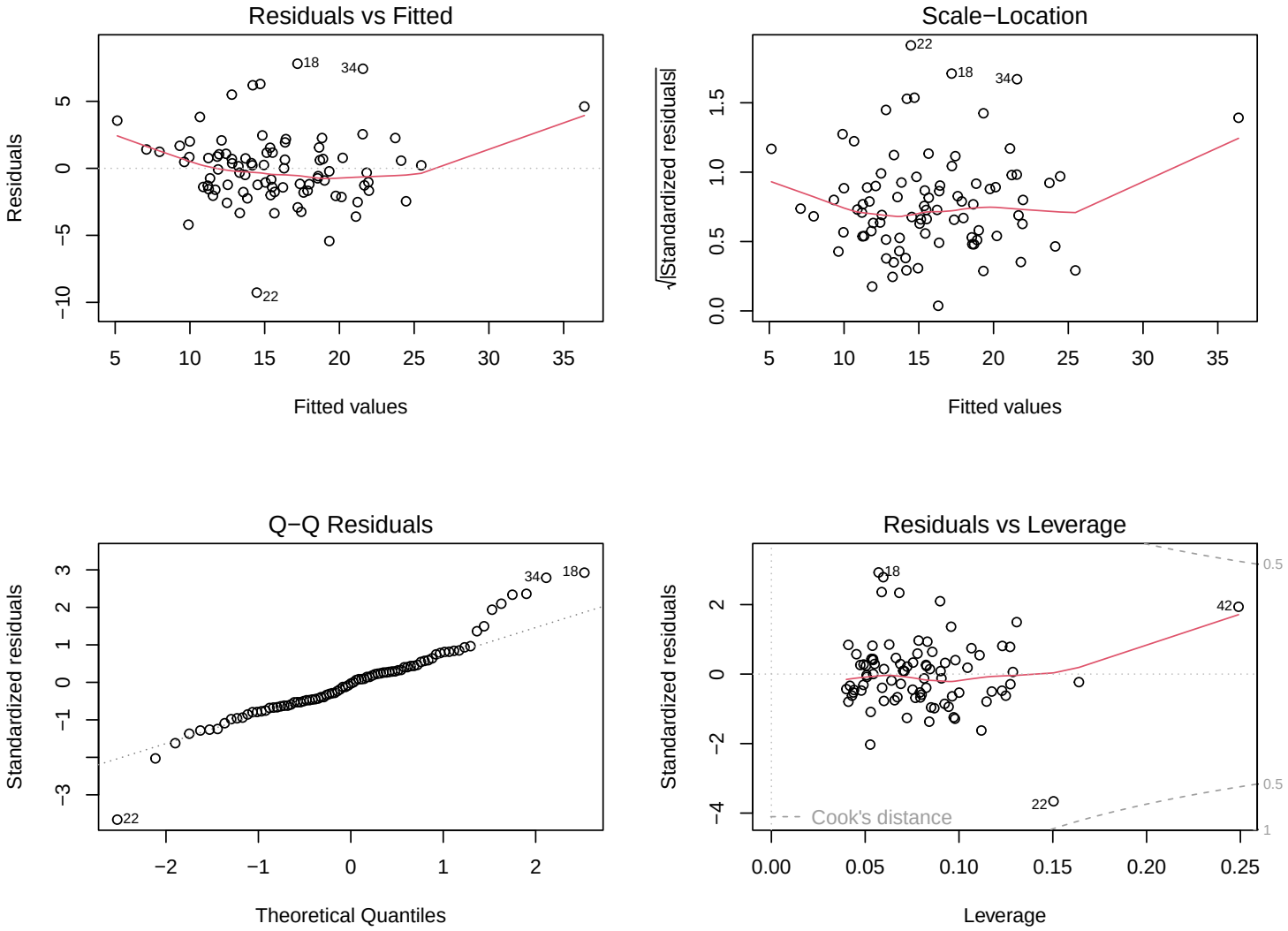
Output 19: Modelo de regressão linear múltiplo após a remoção das variáveis não significativas.

Diagnósticos

Note no **Output 20** que as suposições de linearidade dos parâmetros, homocedasticidade dos resíduos e não correlação dos resíduos parecem estar sendo atendidas. Por outro lado, a suposição de normalidade claramente ainda está sendo violada.

```
# Layout
layout(matrix(c(1, 2, 3, 4), 2, 2))

# Gráficos de diagnóstico
plot(modelo_1)
```



Output 20: Gráficos de diagnósticos do modelo após a remoção das variáveis não significativas: **Residuals vs Fitted** para verificar linearidade; **Scale-Location** para verificar homocedasticidade; **Q-Q Residuals** para verificar normalidade e **Residuals vs Leverage** para verificar pontos influentes.

No **Output 21**, é confirmado que, em geral, o modelo atende aos pressupostos. No entanto, apenas o teste de Kolmogorov-Smirnov aponta normalidade dos resíduos, o que é preocupante. Além disso, note que a observação 22 é de fato um outlier, mas a distância de cook deixa a dúvida sobre a influência desse ponto, pois anteriormente ultrapassava a distância de 0.5 e após a remoção das variáveis não significativas ele ficou abaixo desse valor, porém próximo. É possível que esta observação esteja atrapalhando modelo.

```
# Testando a normalidade dos resíduos
test_press(modelo_1)
```

```
##                                             Estatística        p
## Exact one-sample Kolmogorov-Smirnov test        0.09201 0.4273652
## Lilliefors (Kolmogorov-Smirnov) normality test  0.09936 0.0335903
## Cramer-von Mises normality test                 0.22363 0.0026285
## Shapiro-Wilk normality test                     0.93469 0.0002756
```

```
## Shapiro-Francia normality test                    0.92516 0.0002046
## Anderson-Darling normality test                   1.55992 0.0004783
```

```
# Teste de Breusch-Pagan (Homocedasticidade)
lmtest::bptest(modelo_1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_1
## BP = 11.747, df = 6, p-value = 0.06785
```

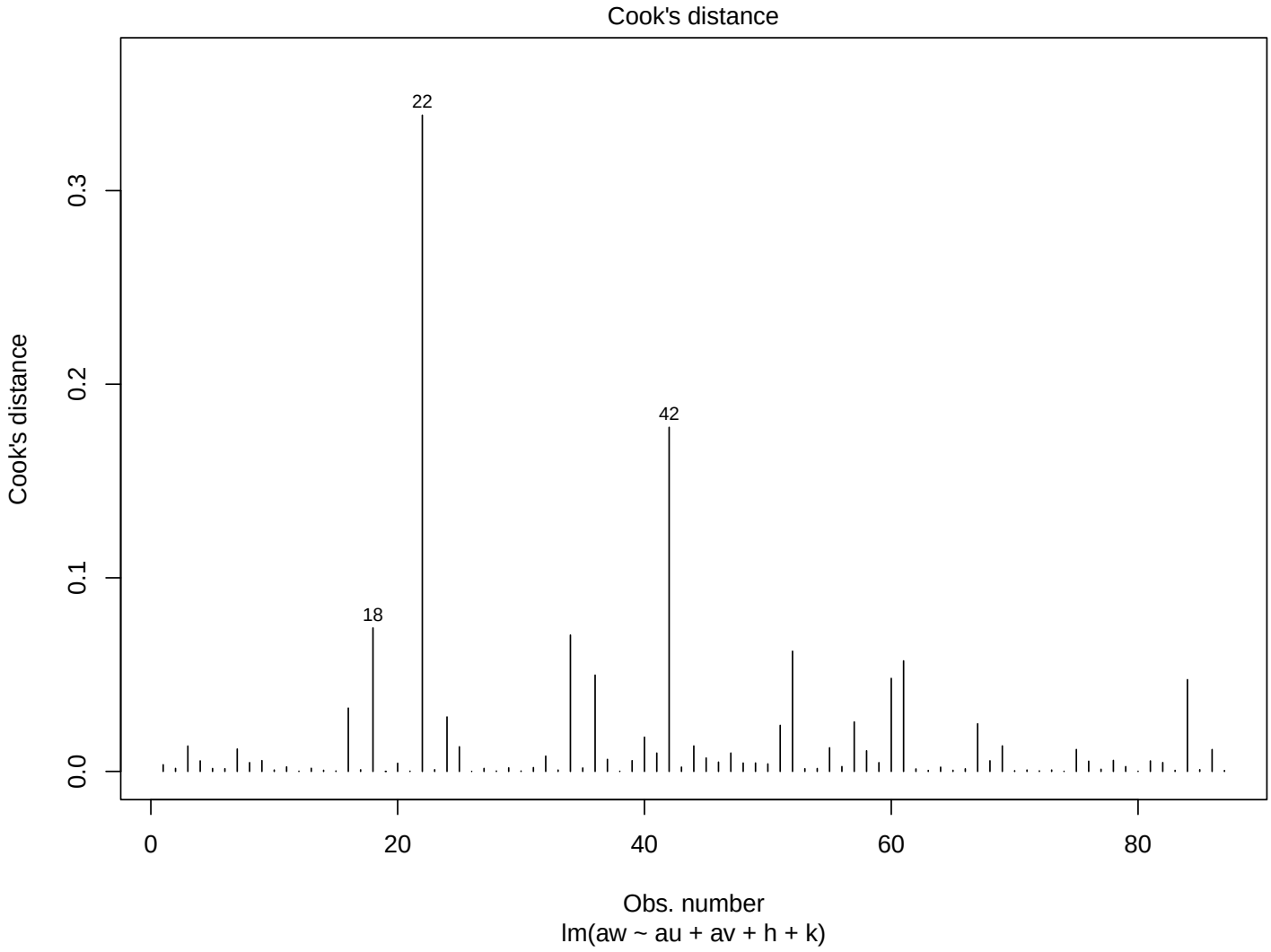
```
# Teste de Durbin-Watson (correlação dos resíduos)
durbinWatsonTest(modelo_1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02583714      1.944411   0.672
##  Alternative hypothesis: rho != 0
```

```
# Verificando outliers
car::outlierTest(modelo_1)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 22 -3.987555         0.00014802     0.012878
```

```
# Observações influentes
cutoff <- 4/((nrow(dados_1)-length(modelo_1$coefficients)-2))
plot(modelo_1, which=4, cook.levels=cutoff)
```



Output 21: Testes estatísticos de disgnósticos do modelo.

Em cenários em que há a presença de outliers e pontos influentes em uma regressão linear, não podemos simplesmente remover esse dado pois pode impactar na característica geral dos dados. Uma forma de lidar com isso é aplicando métodos de regressão robusta.

Para não precisar realizar toda uma nova análise utilizando métodos de regressão robusta, no **Output 22** optei em construir os intervalos de confiança das estimativas utilizando um método de bootstrap chamado de *bias-corrected and accelerated (BCa)*, o qual faz correções de viés e assimetria na distribuição das estimativas bootstrap (ideal em cenários de não normalidade).

```
#sample(1000, 1) = 953
set.seed(953)

modelo_1_boot <- car::Boot(modelo_1, R=1000)
summary(modelo_1_boot)
```

```
##
## Number of bootstrap replications R = 1000
##              original    bootBias   bootSE   bootMed
## (Intercept) 11.129952  3.6970e-01 3.379277 11.509640
```

```
## au           0.118612 -7.6124e-05 0.011789  0.118842
## av           0.043592 -1.6337e-03 0.017521  0.042204
## h           -0.146408 -1.9331e-03 0.030637 -0.147544
## kclasse2    -1.293701 -1.1699e-02 0.821077 -1.309026
## kclasse3    -2.332714  2.8010e-02 0.868703 -2.295738
## kclasse4    -2.296799 -1.6947e-02 0.886089 -2.331156
```

```r
# Colunas com os intervalos de confiança de 95%
intervalos_modelo_1 <- confint(modelo_1_boot)

# Coluna de coeficientes estimados
coefficients_modelo_1 <- as.data.frame(coef(modelo_1))

aux_tabel <- data.frame("Coeficientes" = coefficients_modelo_1,
                        intervalos_modelo_1)
names(aux_tabel) <- c("Coeficientes", "2.5 %", "97.5 %")
kable(aux_tabel, format = "latex")
```

	Coeficientes	2.5 %	97.5 %
(Intercept)	11.1299516	3.1201730	16.5907608
au	0.1186116	0.0946775	0.1398921
av	0.0435916	0.0151985	0.0851323
h	-0.1464081	-0.2064645	-0.0871218
kclasse2	-1.2937006	-2.8922375	0.3076621
kclasse3	-2.3327142	-4.3722679	-0.8825767
kclasse4	-2.2967989	-3.9789173	-0.5317903

Output 22: Estimativas dos coeficientes e intervalos de confiança de 95% obtidos por bootstrap.

4.g

Considerando o modelo ajustado, no cenário extremo em que todas as covariáveis são nulas, o consumo esperado de oxigênio dos pacientes com insuficiência cardíaca no pico do exercício é de 11.13 (unidade: ml/(kg.min); IC(95%) = [3.12, 16.59]).

Além disso, ainda considerando o efeito esperado no consumo de oxigênio: a cada unidade de incremento na carga na esteira ergométrica, é esperado um aumento de 0.12 (unidade: ml/(kg.min); IC(95%) = [0.09, 0.14]). Com cada unidade de incremento na frequência cardíaca (bpm), é esperado um aumento de 0.04 (unidade: ml/(kg.min); IC(95%) = [0.02, 0.09]). A cada unidade de incremento no peso (kg), é esperado um decréscimo de 0.15 (unidade: ml/(kg.min); IC(95%) = [0.09, 0.21]).

Por fim, em comparação com a classe 1 do critério NYHA, é esperado um decréscimo vindo da classe 3 de 2.33 (unidade: ml/(kg.min); IC(95%) = [0.88, 4.37]) no consumo de oxigênio. Em relação à classe 4 do critério NYHA, é esperado um decréscimo de 2.30 (unidade: ml/(kg.min); IC(95%) = [0.53, 3.98]). Comparado com a classe 2 do critério NYHA, é esperado um decréscimo de 1.29. No entanto, neste caso, a classe 2 não foi significativa no modelo ajustado.

Questão 5 (Morettin & Singer)

Modelo a ser considerado:

$$\log\left\{\frac{\pi_i\left(x_i, w_i\right)}{1-\pi_i\left(x_i, w_i\right)}\right\} = \alpha + \beta x_i + \gamma\left(w_i - 5\right), i = 1, \ldots, 50$$

5.a

O parâmetro α é o logaritmo da chance quando todas as covariáveis são nulas (caso mais extremo). Assim, quando o gênero for feminino ($x = 0$) e a idade for 5 ($w = 5$), então, o log da chance de preferência pelo refrigerante Kcola será 0.69. Além disso, ao

aplicar a função exponencial no coeficiente estimado, teremos a interpretação em termos da razão de chances, logo, a chance de preferir Kcola é 99% ($\exp(0.69) = 1.99$) maior do que não preferir.

Considerando o parâmetro $\beta = 0.33$ e ficado a idade, a chance de preferir Kcola sendo do gênero masculino é 39% ($\exp(0.33) = 1.39$) maior do que sendo do gênero feminino. Por fim, considernado o parâmetro $\gamma = -0.03$, cada incremento de 1 ano na idade está associado com uma redução de 3% ($1 - \exp(-0.03) = 1 - 0.97 = 0.03$) na chance de preferência por Kcola.

5.b

$$OR = \frac{\exp[\alpha + \beta x + \gamma\,(w-5)]}{\exp[\alpha + \beta x + \gamma\,(w-5)]} = \frac{\exp[0.69 + 0.33x - 0.03\,(10-5)]}{\exp[0.69 + 0.33x - 0.03\,(15-5)]} =$$

$$= \frac{\exp[-0.03\,(10-5)]}{\exp[0.03\,(15-5)]} = \frac{\exp(-0.15)}{\exp(0.3)} = \exp(0.15) = 1.1618$$

Considerando crianças do mesmo gênero, aquelas que possuem 10 anos de idade possuem 16.18% mais chances de ter preferência por Kcola do que as crianças com 15 anos.

5.c

Para construir o intervalo de confiança para $\exp(\beta)$ pordemos fazer uso da propriedade de invariância, onde, considerando o intervalo para o parâmetro sendo $\beta \pm z^* SE\,(\beta)$, teremos que o intervalo de $\exp(\beta)$ será $\exp(\beta \pm z^* SE\,(\beta))$. Dessa forma, o intervalo de 95% para $\exp(\beta)$ será $\exp(0.33 \pm 1.96 * 0.10) = [1.14, 1.69]$ e para $\exp(\gamma)$ será $\exp(-0.03 \pm 1.96 * 0.005) = [0.96, 0.98]$.

A interpretação pode ser dada da seguinte forma:

Para β: É esperado um aumento entre 14% a 69% na chance de preferência por Kcola em crianças do gênero masculino comparado com as do gênero feminino.

Para γ: Para cada incremento de 1 ano na idade, é esperado uma redução de 2% a 4% na chance de preferência por Kcola.

5.d

$$\log\,(ODDS) = \alpha + \beta x_i + \gamma(w_i - 5), i = 1, \ldots, 50$$

Considerando crianças do gênero masculino com 15 anos:

$$= 0.69 + 0.33 - 0.03(15-5) = 0.72$$

Obtendo o termo em probabilidade:

$$\pi_i = \frac{\exp[\log\,(ODDS)]}{1 + \exp[\log\,(ODDS)]} = \frac{\exp[0.72]}{1 + \exp[0.72]} = 0.6726$$

Assim, a probabilidade de crianças do gênero masculino com 15 anos de idade preferirem Kcola é de 0.6726.

Questão 6 (Morettin & Singer)

Expressão (6.29 do livro):

$$\log\left[\frac{P\,(Y_i = 1 \mid X = x)}{P\,(Y_i = 0 \mid X = x)}\right] = \alpha + \beta x_i, i = 1, \ldots, n$$

Expressão (6.30 do livro):

$$P\left(Y_i = 1 \mid X = x\right) = \frac{\exp\left(\alpha + \beta x_i\right)}{1 + \exp\left(\alpha + \beta x_i\right)}, i = 1, \ldots, n$$

Ambas são equivalentes, irei mostrar a seguir.

$$\log\left[\frac{P\left(Y_i = 1 \mid X = x\right)}{P\left(Y_i = 0 \mid X = x\right)}\right] = \alpha + \beta x_i, i = 1, \ldots, n$$

$$\Leftrightarrow \exp\left[\log\left(\frac{P\left(Y_i = 1 \mid X = x\right)}{P\left(Y_i = 0 \mid X = x\right)}\right)\right] = \exp\left(\alpha + \beta x_i\right), i = 1, \ldots, n$$

$$\Leftrightarrow \frac{P\left(Y_i = 1 \mid X = x\right)}{P\left(Y_i = 0 \mid X = x\right)} = \exp\left(\alpha + \beta x_i\right), i = 1, \ldots, n$$

$$\Leftrightarrow \frac{P\left(Y_i = 1 \mid X = x\right)}{1 - P\left(Y_i = 1 \mid X = x\right)} = \exp\left(\alpha + \beta x_i\right), i = 1, \ldots, n$$

$$\Leftrightarrow P\left(Y_i = 1 \mid X = x\right) = \exp\left(\alpha + \beta x_i\right) - P\left(Y_i = 1 \mid X = x\right)\exp\left(\alpha + \beta x_i\right), i = 1, \ldots, n$$

$$\Leftrightarrow \exp\left(\alpha + \beta x_i\right) = P\left(Y_i = 1 \mid X = x\right) + P\left(Y_i = 1 \mid X = x\right)\exp\left(\alpha + \beta x_i\right), i = 1, \ldots, n$$

$$\Leftrightarrow \exp\left(\alpha + \beta x_i\right) = P\left(Y_i = 1 \mid X = x\right)\left(1 + \exp\left(\alpha + \beta x_i\right)\right), i = 1, \ldots, n$$

$$\Leftrightarrow P\left(Y_i = 1 \mid X = x\right) = \frac{\exp\left(\alpha + \beta x_i\right)}{1 + \exp\left(\alpha + \beta x_i\right)}, i = 1, \ldots, n$$

Note que a expressão de $P\left(Y_i = 1 \mid X = x\right)$ possui o denominador igual ao numerador porém somando 1. Como a função exponencial é sempre positiva para qualquer valor de x, o denominador será sempre uma unidade maior que o numerador para qualquer valor de α, β e x, ou seja, a expressão sempre resultará em um valor entre 0 e 1.

Questão 7 (Morettin & Singer)

A chance (odds) é definida da seguinte forma:

$$ODDS = \frac{P\left(Y_i = 1 \mid X = x\right)}{P\left(Y_i = 0 \mid X = x\right)} = \exp(\alpha + \beta x_i), i = 1, \ldots, n$$

Logo, ao aplicar a função logarítmica em ambos os lado obtemos:

$$\Leftrightarrow \log\left(ODDS\right) = \log\left(\frac{P\left(Y_i = 1 \mid X = x\right)}{P\left(Y_i = 0 \mid X = x\right)}\right) = \alpha + \beta x_i, i = 1, \ldots, n$$

Esse é o modelo a ser considerado. Agora, incluindo o cenário com incremento de uma unidade na variável explicativa podemos construir a OR da seguinte forma:

$$OR = \frac{ODDS^*}{ODDS} = \frac{\exp[\alpha + \beta(x_i + 1)]}{\exp(\alpha + \beta x_i)}, i = 1, \ldots, n$$

Aplicando o logaritmo em ambos os lados:

$$\Leftrightarrow \log(OR) = \log\left[\frac{\exp[\alpha + \beta(x_i + 1)]}{\exp(\alpha + \beta x_i)}\right], i = 1, \ldots, n$$

$$= \log[\exp(\alpha + \beta(x_i + 1))] - \log\left[\exp(\alpha + \beta x_i)\right], i = 1, \ldots, n$$

$$= \alpha + \beta(x_i + 1) - \alpha - \beta x_i, i = 1, \ldots, n$$

$$= \beta x_i + \beta - \beta x_i, i = 1, \ldots, n$$

$$= \beta$$

Questão 8 (Extra)

Quando temos como objetivo modelar algum fenômeno com o intuito de obter previsões, separamos os dados em amostra de treino e amostra de teste, onde na amostra de treino nós treinamos nosso modelo e, em seguida, utilizamos a amostra de teste para verificar sua capacidade de previsão.

Em alguns casos, pode acontecer do modelo ser muito bom e trazer métricas espetaculares, mas na hora de prever novos valores que estão fora da amostra, ele possui um desempenho muito ruim. Quando isso acontece, dizemos que temos um sobreajuste (Overfitting), o que indica que o modelo não tem capacidade de generalização. Por outro lado, podemos obter modelos que simplesmente possuem um ajuste muito pobre, não sendo capazes de captar a variabilidade dos dados corretamente, indicando um subajuste (Underfitting).

Uma forma de lidar com esse problema é dividir a amostra em várias partições de treino e teste, como por exemplo, utilizando métodos de validação cruzada, em que as partições de treino e teste da amostra são feitas diversas vezes para avaliar se aquele modelo de fato é razoável. Assim, sempre precisamos buscar modelos que estejam no meio-termo entre sobreajuste e subajuste e, sempre que possível, aplicar técnicas de validação cruzada ou qualquer outra nesse sentido para garantir que o modelo sendo feito é de fato útil na prática.

Questão 9 (Extra)

Demonstração da expressão do trade-off entre viés e veriância.

Irei partir da seguinte expressão:

$$EQM = E\left[Y_0 - \widehat{f}\left(\mathbf{x}_0\right)\right]^2$$

Considere que $Y_0 = f(\mathbf{x_0}) + \epsilon_0$:

$$= E\left[\left(f\left(\mathbf{x_0}\right) + \epsilon_0 - \hat{f}\left(\mathbf{x_0}\right)\right)^2\right]$$

$$= E\left[\left(f\left(\mathbf{x_0}\right) - \hat{f}\left(\mathbf{x_0}\right)\right)^2\right] + E\left[2\epsilon_0\left(f\left(\mathbf{x_0}\right) - \hat{f}\left(\mathbf{x_0}\right)\right)\right] + E\left[\epsilon_0^2\right]$$

O termo $E\left[2\epsilon_0\left(f\left(\mathbf{x_0}\right) - \hat{f}\left(\mathbf{x_0}\right)\right)\right]$ é zero pois $E\left[\epsilon_0\right] = 0$, logo:

$$= E\left[\left(f\left(\mathbf{x_0}\right) - \hat{f}\left(\mathbf{x_0}\right)\right)^2\right] + E\left[\epsilon_0^2\right]$$

Visto que $Var(\epsilon_0) = E\left[\epsilon_0^2\right] + \left[E\left(\epsilon_0\right)\right]^2$ e $E\left[\epsilon_0\right] = 0$, temos que:

$$= \text{Var}\left[\widehat{f}\left(\mathbf{x}_0\right)\right] + \text{Vies}\left[\widehat{f}\left(\mathbf{x}_0\right)\right]^2 + \text{Var}\left(\epsilon_0\right)$$

Questão 10 (Extra)

Em Estatística e Ciência de Dados, a ideia de flexibilidade também pode ser vista como complexidade, isso porque muitas vezes queremos ser capazes de explicar muito bem um fenômeno através de uma modelagem orientada a dados, por exemplo. A escolha da técnica é crucial para determinar se seremos capazes de obter bons resultados e, ao mesmo tempo, explicá-los. Técnicas mais flexíveis tendem a retornar resultados muito bons se devidamente utilizadas, mas, em contrapartida, tendem a ser mais complexas de serem empregadas, limitando principalmente no quesito de interpretabilidade do fenômeno em questão. Por outro lado, técnicas menos flexíveis, ou seja, com mais limitações seja estruturais ou no quesito de pressupostos e suposições necessárias para que ela possa ser utilizada, tendem a ser mais interpretáveis.

Uma forma de pensar sobre isso é lembrando de modelos de regressão, por exemplo. São modelos com boa interpretabilidade, mas que requerem diversas condições para serem utilizados e validados. Enquanto que modelos mais complexos, como redes neurais, por exemplo, de fato costumam apresentar resultados muito bons, mas são muito mais difíceis de serem interpretados.

No fim, em cenários em que apenas previsões são o suficiente, e que o custo computacional não seja um problema, técnicas com alta flexibilidade possuem muito espaço. No entanto, caso seja um cenário em que a interpretabilidade seja muito importante, modelos menos flexíveis tendem a ser uma melhor opção.

Questão 11 (Extra)

Todo Estatístico ou Cientista de Dados, antes de realizar uma modelagem, precisa se perguntar com qual objetivo em mente o fenômeno deverá ser modelado. Se queremos explicar o fenômeno com base nas relações entre as variáveis independentes e ele, ou se queremos encontrar uma estrutura ou modelo que permita prever valores desse fenômeno que estão fora da amostra em mãos. Em outras palavras, se temos objetivo inferencial, preditivo ou ambos.

A escolha do objetivo é de suma importância, pois a forma com que a modelagem é feita difere. Quando queremos um modelo explicativo (objetivo inferencial), as variáveis independentes no modelo precisam fazer sentido prático para explicar o fenômeno de interesse (desfecho), mesmo que o poder preditivo desse modelo não seja o melhor. Por outro lado, quando queremos um modelo com alta capacidade preditiva, as variáveis independentes incluídas no modelo não necessariamente precisam ter sentido prático, desde que aumentem o poder de previsão do modelo, mesmo que não seja tão fácil interpretar os parâmetros do modelo posteriormente.

Note que esses objetivos possuem pontos positivos e negativos, onde no objetivo inferencial temos maior capacidade explicativa e menor capacidade preditiva, e no objetivo preditivo temos maior capacidade preditiva e menor capacidade explicativa. Assim, muitas vezes queremos um modelo que tenha alta capacidade preditiva, mas ao mesmo tempo queremos a possibilidade de interpretar seus parâmetros e resultados das relações entre desfecho e variáveis independentes com maior facilidade, juntando assim ambos os objetivos e criando um modelo intermediário que seja razoável em ambos os cenários. Um exemplo disso é a regressão logística, que suporta a capacidade preditiva e ao mesmo tempo possui alta interpretação dos parâmetros, inclusive fazendo uma ligação direta com quantidades como a razão de chances, por exemplo.

Esse tema também tem uma certa ligação com a ideia de flexibilidade de modelos, visto que quanto mais flexíveis eles tendem a ser, mais complexos são, e consequentemente a dificuldade de interpretação tende a aumentar, o que pode ser um problema caso o objetivo seja inferencial ou um objetivo híbrido. Por outro lado, em um cenário cujo objetivo seja estritamente preditivo, modelos altamente flexíveis podem não ser um problema, com algumas ressalvas a serem consideradas a respeito da capacidade computacional necessária para serem executados.

No fim, é interessante sempre que possível buscar um equilíbrio entre a capacidade de interpretação e a capacidade preditiva ao modelar fenômenos, pois, embora queiramos prever, na maioria das vezes é de suma importância também entender e interpretar o fenômeno corretamente para trazer maior confiança nas previsões sendo realizadas.

Questão 12 (Extra)

Quando escolhemos o objetivo de predição ao modelar um determinado fenômeno, estamos principalmente interessados na capacidade de previsão do modelo, quanto maior, melhor. Assim, por não haver uma preocupação primária na interpretação do modelo, é comum que o modelo escolhido para previsão, que contenha a melhor capacidade de previsão, seja um modelo caixa preta, ou seja, um modelo não interpretável. Dessa forma, quando queremos certo poder de interpretação nesses cenários, podemos recorrer a métodos de interpretabilidade que podem ser aplicados a qualquer modelo preditivo previamente ajustado.

A Ornella em seu trabalho de iniciação científica discute esses métodos de interpretabilidade, entre eles temos o seguinte:

1 - **Gráfico de dependência parcial**: Considerando um grupo de covariáveis de interesse e um grupo das demais covariáveis, o gráfico de dependência parcial se baseia em uma função chamada de função de dependência parcial, a qual calcula o efeito médio das covariáveis de interesse ao marginalizar a distribuição das previsões sobre as demais covariáveis (nas quais não há interesse em explicar), tornando a função dependente apenas das covariáveis de interesse. A interpretação possibilitada a partir deste método só é válida no cenário em que as covariáveis de interesse são independentes das demais covariáveis.

2 - **Gráfico da esperança condicional individual**: Neste método, o grau de dependência entre variável e previsão é considerado para cada dado da amostra. Possui uma relação com o método da dependência parcial, onde o gráfico da dependência parcial é a média das curvas do gráfico da esperança condicional individual, e também está sujeito à mesma limitação, em que apenas é válido quando as covariáveis de interesse são independentes das demais covariáveis.

3 - **Gráfico de efeitos locais acumulados**: Tem por objetivo calcular o efeito médio das covariáveis sobre as previsões do modelo. Ou seja, possui o mesmo objetivo do gráfico de dependência parcial, porém, faz uso da distribuição condicional das covariáveis. É uma alternativa ao gráfico de dependência parcial e ao gráfico da esperança condicional individual, pois o efeito estimado das covariáveis de interesse não é interferido por variáveis correlacionadas, o que burla de certa forma a limitação desses dois últimos métodos.

4 - **Interação das covariáveis**: Avalia o peso da interação de uma determinada covariável com as demais, chamada de interação bidirecional, e também a interação total, ao avaliar todos os pares possíveis de covariáveis. A força dessa interação pode ser medida através do quanto a variação da previsão depende da interação sendo avaliada. Sua limitação é que funciona bem apenas para variáveis independentes.

5 - **Modelo interpretável substituto global**: É um método em que um modelo substituto é treinado com o intuito de aproximar as previsões de um modelo não interpretável. O modelo substituto usa a previsão do modelo não interpretável como desfecho, assim, podemos utilizá-lo para tirar conclusões sobre o modelo de interesse interpretando o modelo substituto.

6 - **Modelo interpretável substituto global e modelo interpretável substituto local**: O método global é um método em que um modelo substituto é treinado com o intuito de aproximar as previsões de um modelo não interpretável. O modelo substituto usa a previsão do modelo não interpretável como desfecho, assim, podemos utilizá-lo para tirar conclusões sobre o modelo de interesse interpretando o modelo substituto. Além disso, o método local tem a mesma ideia, porém, a ideia se concentra em explicar as previsões de forma individual em vez de global.

7 - **Valores shapley**: É um método baseado na teoria dos jogos, que visa distribuir de forma justa as previsões às covariáveis de um subconjunto de variáveis de acordo com a contribuição (positiva ou negativa) individual de cada uma no modelo preditivo em relação à previsão média feita para todos os dados. É um dos métodos mais completos com intuito de explicabilidade, porém com um custo computacional maior.

Note que todos possuem cenários nos quais eles podem ser utilizados, então é importante verificar para o cenário em questão qual seria o método mais apropriado.

Questão 13 (Extra)