Análise - Vendas de Vestuário

José Ramon Severo Alves Leonardo Almeida Farias

2025-09-24

Índice

1 Introdução

Análise estatística do dataset "moda_vestuario_vendas.xlsx", que contém diversas observações sobre vendas de peças de roupas em diferentes períodos de tempo e em diferentes contextos, incluindo dados categóricos e númericos que podem influenciar esse valor.

O principal objetivo deste trabalho é identificar que fatores influenciam o volume de vendas de peças de roupas, através da análise exploratória de dados e da regressão linear simples.

2 Dados da Análise

```
dados <- read_excel("moda_vestuario_vendas.xlsx")
glimpse(dados)
```

```
Rows: 1,200
Columns: 13
$ Vendas                  <dbl> 1335, 1911, 5231, 3297, 3886, 3424, 1790, 2
$ Investimento_marketing  <dbl> 55.27, 70.81, 112.20, 84.85, 103.01, 94.43,
$ Alcance_midias_sociais  <dbl> 509.9, 624.6, 1065.6, 751.8, 895.2, 787.3,
$ Preco_medio             <dbl> 118.34, 96.96, 96.31, 172.66, 165.07, 108.6
$ Desconto_medio          <dbl> 3.24, 5.66, 11.18, 10.38, 10.07, 11.72, 12.
$ N_skus_ativos           <dbl> 153, 349, 488, 608, 358, 196, 470, 331, 399
$ Estoque_medio           <dbl> 18.80, 38.92, 40.88, 32.50, 53.21, 35.59, 3
$ Satisfacao_clientes     <dbl> 9.05, 8.33, 10.00, 8.42, 7.10, 6.61, 5.90,
$ Taxa_devolucao          <dbl> 3.73, 5.61, 2.89, 0.76, 6.34, 5.43, 7.57, 6
$ Trafego_site            <dbl> 206.7, 363.1, 242.5, 373.8, 350.3, 237.0, 3
$ Taxa_conversao_online   <dbl> 2.68, 3.51, 2.93, 2.49, 2.24, 3.75, 0.59, 0
$ Estacao                 <chr> "Inverno", "Primavera", "Verão", "Primavera
$ Campanha_influenciadores <chr> "Sim", "Não", "Sim", "Não", "Sim", "Não", "
```

Nome	Descrição
Vendas	Total de peças vendidas no mês
Investimento_marketing	Gastos com marketing no mês, em milhares de reais

Nome	Descrição
Alcance_midias_sociais	Quantidade de pessoas atingidas nas redes sociais, em milhares de pessoas
Preco_medio	Preço médio das peças, em reais
Desconto_medio	Desconto médio aplicado sobre o preço médio, em porcentagem
N_skus_ativos	Quantidade de SKUs ativas, equivalente ao número de peças diferentes e suas variações em estoque
Estoque_medio	Nível médio de estoque disponível no mês, em milhares de peças
Satisfacao_clientes	Índice de satisfação média do cliente, entre 0-10
Taxa_devolucao	Percentual de pedidos devolvidos ou cancelados
Trafego_site	Visitas ao site no mês, em milhares de visitas
Taxa_conversao_online	Porcentagem de visitantes que efetivaram uma compra online
Estacao	Em qual estação foi feita a análise, outono, inverno, verão ou primavera
Campanha_influenciadores	Se teve ou não campanhas de marketing de influenciadores, booleano.

```
print(skim(dados))
```

```
-- Data Summary ------------------------
                           Values
Name                       dados
Number of rows             1200
Number of columns          13
_________________________
Column type frequency:
  character                2
  numeric                  11
_________________________
Group variables            None

-- Variable type: character -----------------------------------------------
  skim_variable            n_missing complete_rate min max empty n_unique
1 Estacao                          0             1   5   9     0        4
2 Campanha_influenciadores         0             1   3   3     0        2
  whitespace
```

```
1          0
2          0

-- Variable type: numeric -----------------------------------------------
   skim_variable          n_missing complete_rate    mean      sd     p0
 1 Vendas                         0          1    3022.   823.    680    245
 2 Investimento_marketing         0          1      80.2   25.4    -
1.72   64.0
 3 Alcance_midias_sociais         0          1     723.   230.    -
11.6   573.
 4 Preco_medio                   12          0.99  120.    24.7    45.6   10
 5 Desconto_medio                12          0.99   10.1    4.89   -
5.63    6.51
 6 N_skus_ativos                  0          1     354.   118.     30     27
 7 Estoque_medio                 12          0.99   40.0   14.9    -
2.8    29.8
 8 Satisfacao_clientes           12          0.99    7.47   1.45    2.68
 9 Taxa_devolucao                12          0.99    6.10   2.92    0
10 Trafego_site                  0          1     298.    99.4    -
9.3   237.
11 Taxa_conversao_online         0          1       2.77   0.923   0.2
       p50     p75    p100 hist
 1 3005    3581.   5751    ▂▆▆▃▂
 2   80.8    98.4   156.    ▂▅▆▃▂
 3  728.    883.   1402.    ▂▅▆▃▂
 4  119.    137.    199.    ▂▅▆▃▂
 5   10.2    13.4    24.3   ▂▅▆▃▂
 6  355     436     692     ▂▅▆▃▂
 7   39.7    50.0    94.1   ▂▅▆▂▂
 8    7.58    8.53   10     ▂▃▆▆▇
 9    6.14    8.03   15.9   ▂▅▆▂▂
10  295.    357.    636.    ▂▅▆▃▂
11    2.76    3.35    5.97  ▂▅▆▂▂
```

Análisando as 1200 linhas, vemos que os valores numéricos estão entre alcances razoáveis, mas algumas linhas tem valores faltando nas colunas `Preco_medio`, `Desconto_medio`, `Estoque_medio`, `Satisfacao_clientes` e `Trafego_site`.Especificamente:

Coluna	Quantidade de observações faltando
Preco_medio	12
Desconto_medio	12
Estoque_medio	12
Satisfacao_clientes	12
Trafego_site	12

Para manter a integridade desse dataset relativamente pequeno, escolhemos remover essas linhas incompletas, diminuindo o dataset para ter 1140 linhas.

```
dados <- na.omit(dados)
glimpse(dados)
```

```
Rows: 1,140
Columns: 13
$ Vendas                   <dbl> 1335, 1911, 5231, 3297, 3886, 3424, 1790, 2
$ Investimento_marketing   <dbl> 55.27, 70.81, 112.20, 84.85, 103.01, 94.43,
$ Alcance_midias_sociais   <dbl> 509.9, 624.6, 1065.6, 751.8, 895.2, 787.3,
$ Preco_medio              <dbl> 118.34, 96.96, 96.31, 172.66, 165.07, 108.6
$ Desconto_medio           <dbl> 3.24, 5.66, 11.18, 10.38, 10.07, 11.72, 12.
$ N_skus_ativos            <dbl> 153, 349, 488, 608, 358, 196, 470, 331, 399
$ Estoque_medio            <dbl> 18.80, 38.92, 40.88, 32.50, 53.21, 35.59, 3
$ Satisfacao_clientes      <dbl> 9.05, 8.33, 10.00, 8.42, 7.10, 6.61, 5.90,
$ Taxa_devolucao           <dbl> 3.73, 5.61, 2.89, 0.76, 6.34, 5.43, 7.57, 6
$ Trafego_site             <dbl> 206.7, 363.1, 242.5, 373.8, 350.3, 237.0, 3
$ Taxa_conversao_online    <dbl> 2.68, 3.51, 2.93, 2.49, 2.24, 3.75, 0.59, 0
$ Estacao                  <chr> "Inverno", "Primavera", "Verão", "Primavera
$ Campanha_influenciadores <chr> "Sim", "Não", "Sim", "Não", "Sim", "Não", "
```

Nossa interpretação desse dataset é que cada linha é referente a uma diferente loja ou site que teve seu dados coletados em algum mês, ou até mesmo a mesma loja em períodos diferentes.

3 Análises bivariadas

3.1 Entre variáveis numéricas

```r
numeric_vars <- dados %>% select(where(is.numeric))
cor_matrix <- cor(numeric_vars, use = "pairwise.complete.obs")
print(cor_matrix)
```

```
                          Vendas Investimento_marketing
Vendas                1.000000000            0.820761275
Investimento_marketing 0.820761275            1.000000000
Alcance_midias_sociais 0.835558225            0.967589069
Preco_medio            0.030450734            0.046599988
Desconto_medio        -0.006494389           -0.055251250
N_skus_ativos          0.008739960            0.019589234
Estoque_medio         -0.040552317           -0.043702703
Satisfacao_clientes   -0.010366202           -0.049457274
Taxa_devolucao         0.003356714            0.009422445
Trafego_site           0.007744699            0.008776480
Taxa_conversao_online  0.082967410           -0.010961200
                      Alcance_midias_sociais  Preco_medio Desconto_medio
Vendas                           0.8355582251  0.030450734   -
0.006494389
Investimento_marketing           0.9675890686  0.046599988   -
0.055251250
Alcance_midias_sociais           1.0000000000  0.040402527   -
0.058350502
Preco_medio                      0.0404025271  1.000000000    0.025410515
Desconto_medio                  -0.0583505020  0.025410515    1.000000000
N_skus_ativos                    0.0289531496  0.022349857   -
0.032488814
Estoque_medio                   -0.0399791692 -0.019514303   -
0.011636250
Satisfacao_clientes             -0.0537672532  0.005036666   -
0.001297739
Taxa_devolucao                   0.0129218646  0.046121034   -
0.030950986
Trafego_site                     0.0179218532  0.004613062    0.001152242
Taxa_conversao_online            0.0005167121  0.012029476   -
```

```
0.015561417
                      N_skus_ativos Estoque_medio Satisfacao_clientes
Vendas                 0.0087399601  -0.040552317        -
0.010366202
Investimento_marketing  0.0195892341  -0.043702703        -
0.049457274
Alcance_midias_sociais  0.0289531496  -0.039979169        -
0.053767253
Preco_medio             0.0223498568  -0.019514303         0.005036666
Desconto_medio         -0.0324888140  -0.011636250        -
0.001297739
N_skus_ativos           1.0000000000   0.016389485         0.030821169
Estoque_medio           0.0163894854   1.000000000         0.022958312
Satisfacao_clientes     0.0308211690   0.022958312         1.000000000
Taxa_devolucao          0.0523400757   0.006266208        -
0.014859097
Trafego_site            0.0001002371  -0.053861973         0.014467925
Taxa_conversao_online  -0.0179546717  -0.066269365         0.001732701
                      Taxa_devolucao  Trafego_site Taxa_conversao_online
Vendas                   0.003356714  0.0077446985          0.0829674103
Investimento_marketing    0.009422445  0.0087764801         -
0.0109612001
Alcance_midias_sociais    0.012921865  0.0179218532          0.0005167121
Preco_medio               0.046121034  0.0046130617          0.0120294759
Desconto_medio           -0.030950986  0.0011522423         -
0.0155614173
N_skus_ativos             0.052340076  0.0001002371         -
0.0179546717
Estoque_medio             0.006266208 -0.0538619734         -
0.0662693649
Satisfacao_clientes      -0.014859097  0.0144679246          0.0017327008
Taxa_devolucao            1.000000000  0.0024234510         -
0.0023184420
Trafego_site              0.002423451  1.0000000000         -
0.0162387294
Taxa_conversao_online    -0.002318442 -0.0162387294          1.0000000000
```

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
corrplot(cor_matrix, method = "color", addCoef.col = "black")
```

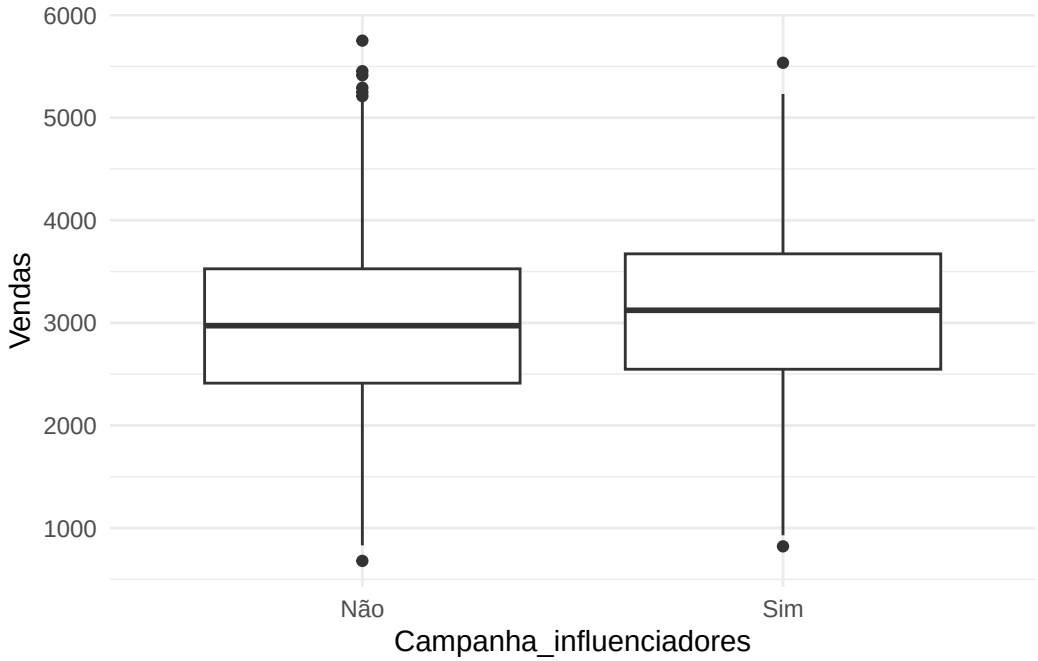
Encontramos algumas corelações entre nossa variável resposta (`Vendas`) e as demais, principalmente entre ela, `Investimento_marketing` e `Alcance_midias_sociais`, com valores de 82% e 84% respectivamente. Surpreendetemente as demais colunas tem correlações **extremamente fracas** com `Vendas`, indicando que não afetam o volume de vendas significativamente.

Além disso, encontramos uma correlação **extremamente forte** entre `Investimento_marketing` e `Alcance_midias_sociais`, de 97%. Isso indica multicolinearidade, então essas variáveis tem quase a mesma informação.

3.2 Variável numérica em função de categórica

3.2.1 Análise entre vendas e a presença de campanha com influenciadores

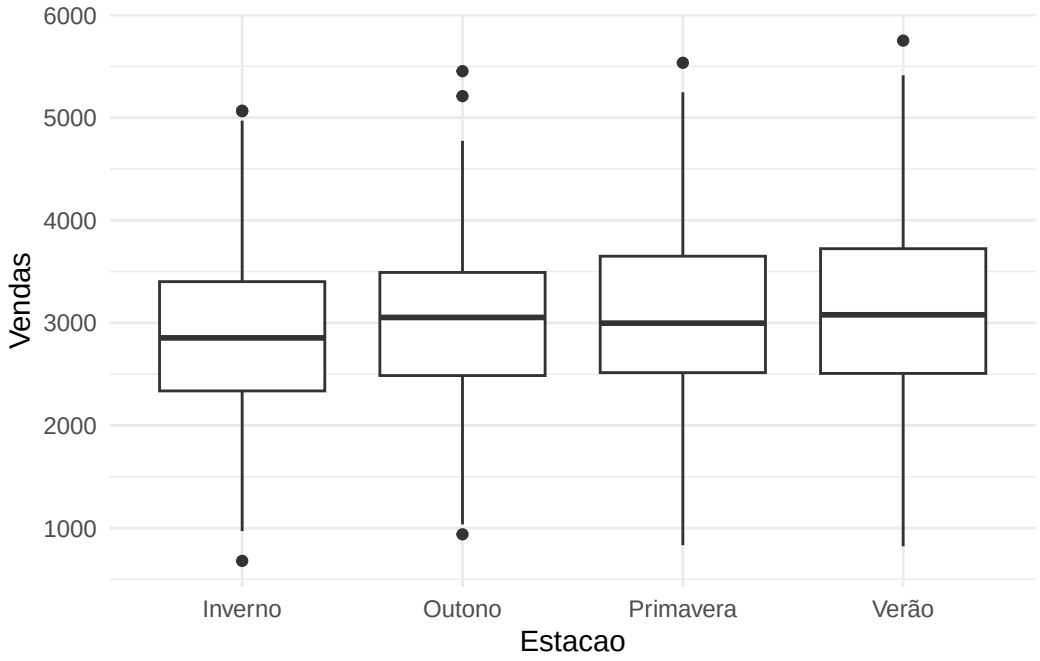
```
ggplot(dados, aes(x = Campanha_influenciadores, y = Vendas)) +
geom_boxplot() + theme_minimal()
```



A análise bivariada entre vendas e presença de campanha com influencers evidencia um leve aumento nas vendas, não tem influência sensível sobre a dispersão dos dados entretanto.

3.2.2 Análise entre vendas e a estação análisada

```
ggplot(dados, aes(x = Estacao, y = Vendas)) +
geom_boxplot() + theme_minimal()
```



Já a análise bivariada entre vendas e estações do ano temos uma grande diferença dados de cada categoria. No inverno tem uma dispersão menor, uma mediana menor, e menos vendas quando comparada com as outras estações.

No outono é perceptível que a mediana está na parte superior dos dados, com uma baixa e desigual dispersão de dados e a quantidade de vendas está ainda menor que a do inverno.

Na primavera, em contraste com o inverno temos uma mediana baixa, uma dispersão de dados maior e também desigual. A quantidade de vendas em comparação com outono e inverno é bem maior.

No verão temos uma grande dispersão entre os dados e uma mediana no meio. A quantidade de vendas em comparação com as estações de outono e inverno são bem maiores.

4 Regressão Linear

4.1 Modelos Lineares

4.1.1 Modelo 1 - Investimento e Alcance

```
modelo1 <- lm(Vendas ~ Investimento_marketing + Alcance_midias_sociais + T
summary(modelo1)
```

```
Call:
lm(formula = Vendas ~ Investimento_marketing + Alcance_midias_sociais +
    Taxa_conversao_online, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-1572.60  -301.74   -25.16   290.87  2268.31

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            635.5038    59.1906  10.737  < 2e-
16 ***
Investimento_marketing   6.7296     2.0583   3.270  0.00111 **
Alcance_midias_sociais   2.2667     0.2272   9.977  < 2e-
16 ***
Taxa_conversao_online   75.3408    14.2458   5.289 1.48e-
07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 445.7 on 1136 degrees of freedom
Multiple R-squared:  0.7077,    Adjusted R-squared:  0.7069
F-statistic: 916.9 on 3 and 1136 DF,  p-value: < 2.2e-16
```

O modelo 1 tem R^2 igual a 0,7, ou seja 70% da variação no volume de vendas é explicado por apenas essas duas variáveis.

```
print(modelo1 %>% tbl_regression() %>% as_tibble())
```

```
# A tibble: 3 x 4
  `**Characteristic**`   `**Beta**` `**95% CI**` `**p-value**`
  <chr>                  <chr>      <chr>        <chr>
1 Investimento_marketing 6.7        2.7, 11      0.001
2 Alcance_midias_sociais 2.3        1.8, 2.7     <0.001
3 Taxa_conversao_online  75         47, 103      <0.001
```

```
vif(modelo1)
```

```
Investimento_marketing Alcance_midias_sociais  Taxa_conversao_online
             15.713379               15.711495                1.002064
```

VIF é uma medida que testa multicolinearidade em modelos, sendo valores acima de 10 considerados problemas sérios [2].

A Taxa_conversao_online apresenta um VIF muito próximo de 1, indicando que não tem multicolinearidade com as outras. Já as duas outras medidas usadas no modelo 1 apresentam VIFs extremamente altos de cerca de 15,68, indicando novamente a grande correlação entre essas variáveis, que pode ser problemática.

Esse resultado confirma o que, intuitivamente, já era de se esperar, que Alcance_midias_sociais e Investimento_marketing fossem relacionadas, já que uma deve levar a outra.

Para simplificar o modelo, evitar redundâncias e aumentar a estabilidade do modelo, vamos análisar qual das variáveis devemos remover.

4.1.2 Modelo 2 - Removendo alcance

```
modelo2 <- lm(formula = Vendas ~ Investimento_marketing + Taxa_conversao_o
summary(modelo2)
```

```
Call:
lm(formula = Vendas ~ Investimento_marketing + Taxa_conversao_online,
    data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1604.90  -305.00   -27.26   300.93  2747.12

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           664.0970    61.6301   10.78  < 2e-
16 ***
Investimento_marketing  26.6014     0.5413   49.14  < 2e-
16 ***
Taxa_conversao_online   81.6015    14.8359    5.50 4.68e-
08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 464.6 on 1137 degrees of freedom
Multiple R-squared:  0.6821,    Adjusted R-squared:  0.6815
F-statistic:  1220 on 2 and 1137 DF,  p-value: < 2.2e-16
```

Retirando o alcance temos uma pequena diferença, trazendo um R^2 ligeiramente menor de 0.6815.

```
vif(modelo2)
```

```
Investimento_marketing  Taxa_conversao_online
               1.00012                1.00012
```

O VIF do modelo 2 é significante mente menor que o do modelo 3.

Em `Investimento_marketing` e em `Taxa_conversao_online` agora temos um VIF muito proximo de 1, indicando que eles não tem correlação. Esse pode ser um modelo melhor já que evita usar variáveis que possuem correlação.

4.1.3 Modelo 3 - Removendo investimentos

```
modelo3 <- lm(formula = Vendas ~ Alcance_midias_sociais + Taxa_conversao_o
summary(modelo3)
```

```
Call:
lm(formula = Vendas ~ Alcance_midias_sociais + Taxa_conversao_online,
```

```
    data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-1553.04  -300.66   -21.46   281.84  2857.15

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           661.17829   58.91691  11.222  < 2e-
16 ***
Alcance_midias_sociais   2.98546    0.05756  51.868  < 2e-
16 ***
Taxa_conversao_online   73.22684   14.29163   5.124 3.52e-
07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 447.6 on 1137 degrees of freedom
Multiple R-squared:  0.705, Adjusted R-squared:  0.7045
F-statistic:  1358 on 2 and 1137 DF,  p-value: < 2.2e-16
```

Removendo os investimentos também, temos um R^2 de 0.705.

```
vif(modelo3)
```

```
Alcance_midias_sociais  Taxa_conversao_online
                     1                      1
```

No modelo 3 também os valores de VIF foram ainda menores que os modelos do modelo 2, também muito próximos de 1.

4.2 Escolhendo o melhor modelo

O modelo 3 tem praticamente a mesma significância do modelo 1 (R^2 = 0,7045 no modelo 3 vs R^2 = 0,7069 no modelo 1) que é consideravelmente maior que a do modelo 2 (R^2 = 0,6734).

```
AIC(modelo2, modelo3)
```

```
        df      AIC
modelo2  4 17242.28
modelo3  4 17157.20
```

AIC(Akaike Information Criterion) é uma das principais ferramentas de comparação entre modelos no mesmo dataset, sendo uma "[...] medida surpreendentemente simples da variância média de valores fora da amostra"[1] > O menor AIC é melhor, indicando que o modelo perdeu menos informação em relação ao outro.

Com ela, temos outros indício de que o modelo 3 é o mais adequado para a situação.

```
str (anova(modelo1, modelo3))
```

```
Classes 'anova' and 'data.frame':   2 obs. of  6 variables:
 $ Res.Df   : num  1136 1137
 $ RSS      : num  2.26e+08 2.28e+08
 $ Df       : num  NA -1
 $ Sum of Sq: num  NA -2123869
 $ F        : num  NA 10.7
 $ Pr(>F)   : num  NA 0.00111
 - attr(*, "heading")= chr [1:2] "Analysis of Variance Table\n" "Model 1: Ve
```

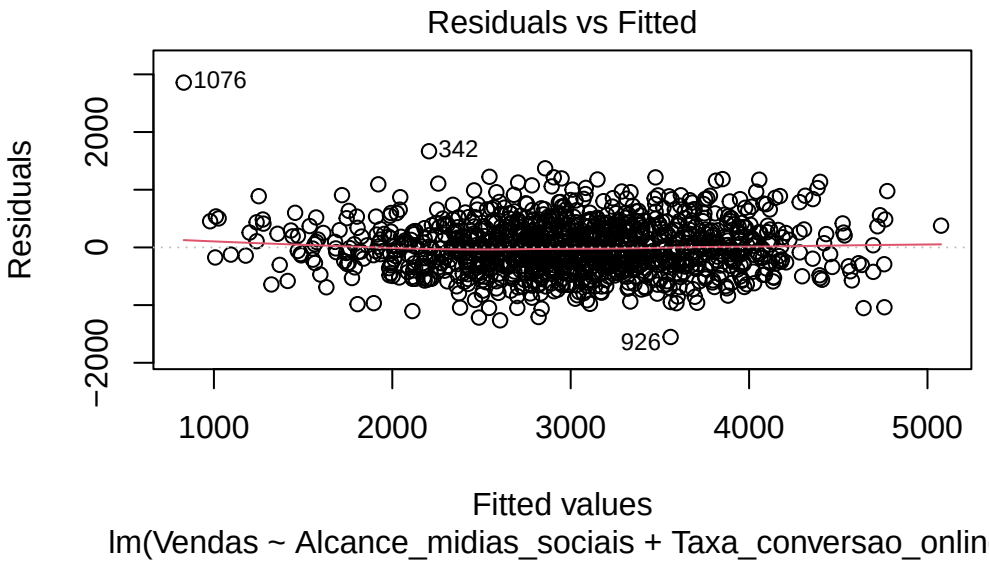
Em contraponto, o teste ANOVA resultou em um p muito menor que 5%, indicando que o modelo 1(mais complexo) tem uma capacidade maior de explicar as vendas em comparação ao modelo 3, como vimos anteriormente com a comparação de seus R^2.

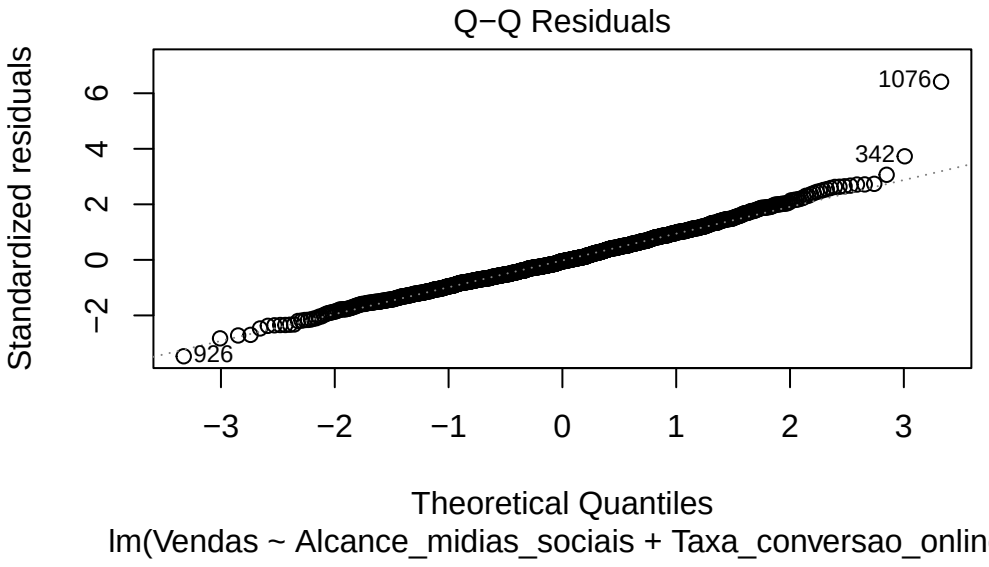
Entretanto, o ANOVA testa apenas se o modelo é melhor para explicar a variável resposta do que utilizar a média, não considerando os possíveis malefícios de usar um modelo mais complexo pode trazer.

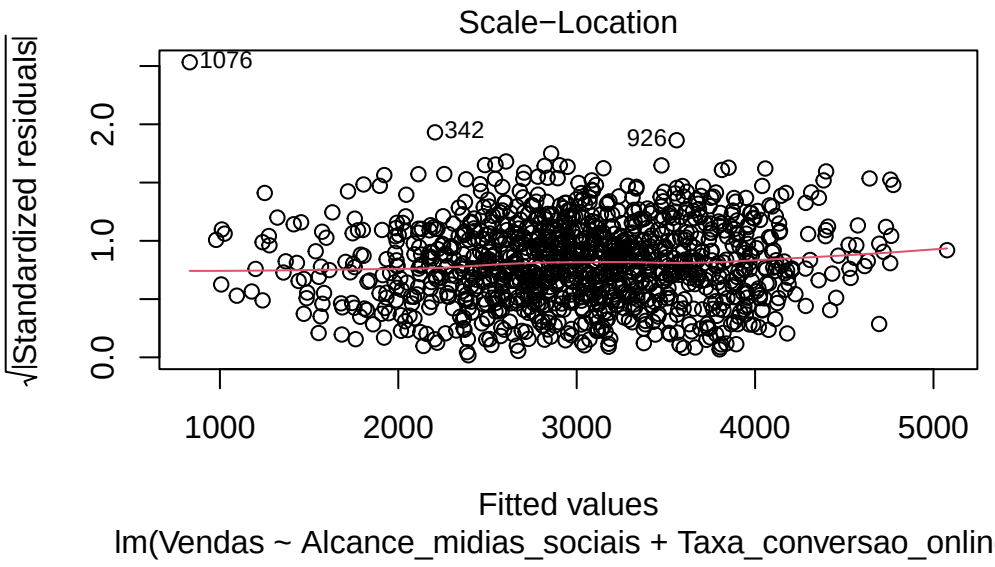
Como visto nos testes anteriores, o modelo 1 ganha pouca melhoria no ajuste(<1%) com esse aumento de complexidade, logo vamos esolher o modelo 3 para continuar nossa análise.

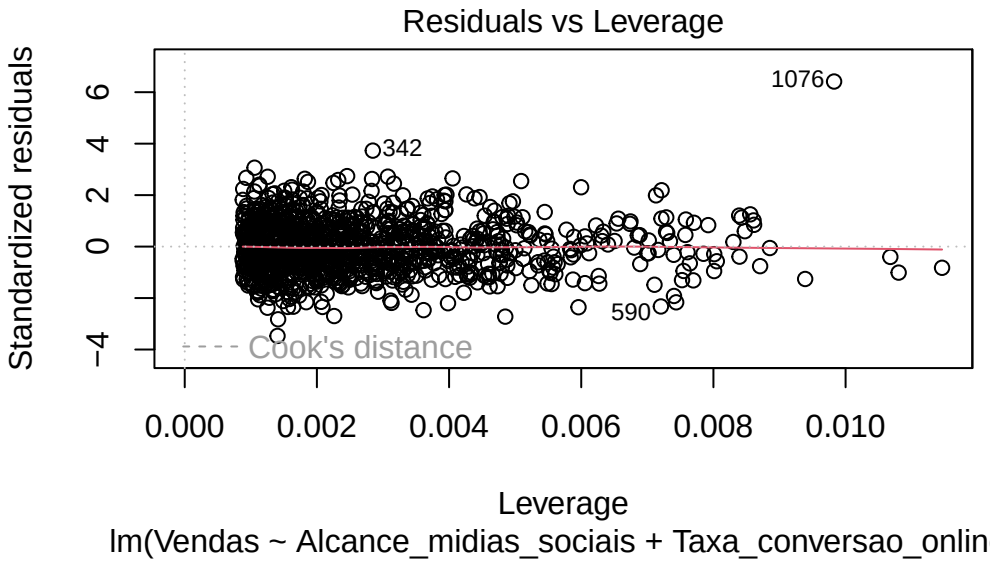
5 Análisando o modelo escolhido

```
plot(modelo3)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Vendas ~ Alcance_midias_sociais + Taxa_conversao_online)
1076
342
926
−2000
0
2000
1000
2000
3000
4000
5000


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Vendas ~ Alcance_midias_sociais + Taxa_conversao_online)
1076
342
926
−2
0
2
4
6
−3
−2
−1
0
1
2
3

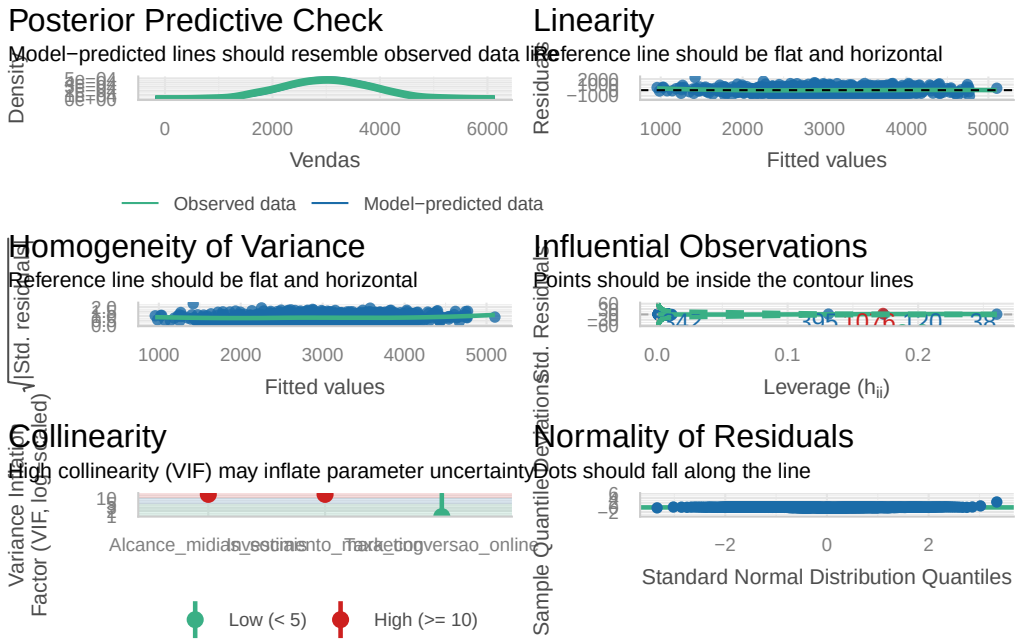

Scale−Location
√|Standardized residuals|
1076
342
926
0.0
1.0
2.0
1000
2000
3000
4000
5000
Fitted values
lm(Vendas ~ Alcance_midias_sociais + Taxa_conversao_online)


Residuals vs Leverage
Standardized residuals
1076
342
590
Cook's distance
−4
0
2
4
6
0.000
0.002
0.004
0.006
0.008
0.010
Leverage
lm(Vendas ~ Alcance_midias_sociais + Taxa_conversao_online)

5.1 Análise dos pressupostos

5.1.1 Do modelo original

```
check_model(modelo1, title_size = 10, base_size = 8)
```



1. Verificação Preditiva Posterior

A curva dos dados preditos (linha verde) está razoavelmente próxima da curva dos dados observados (linha azul).Tendo assim uma boa adequação geral do modelo.

2. Linearidade

Os pontos estão bem distribuídos em torno da linha zero. A linha de tendência verde está quase perfeitamente plana

3. Homocedasticidade

A dispersão dos pontos é bastante uniforme da esquerda para a direita. A linha de tendência verde é praticamente horizontal. Não há o formato de “cone” ou “funil” que indicaria um problema (heterocedasticidade).

4. Normalidade dos Resíduos

Os pontos se alinham quase perfeitamente com a linha reta. Há um pequeno desvio

nos extremos, o que é muito comum e geralmente não é preocupante em conjuntos de dados grandes.

5. Influência de Observações

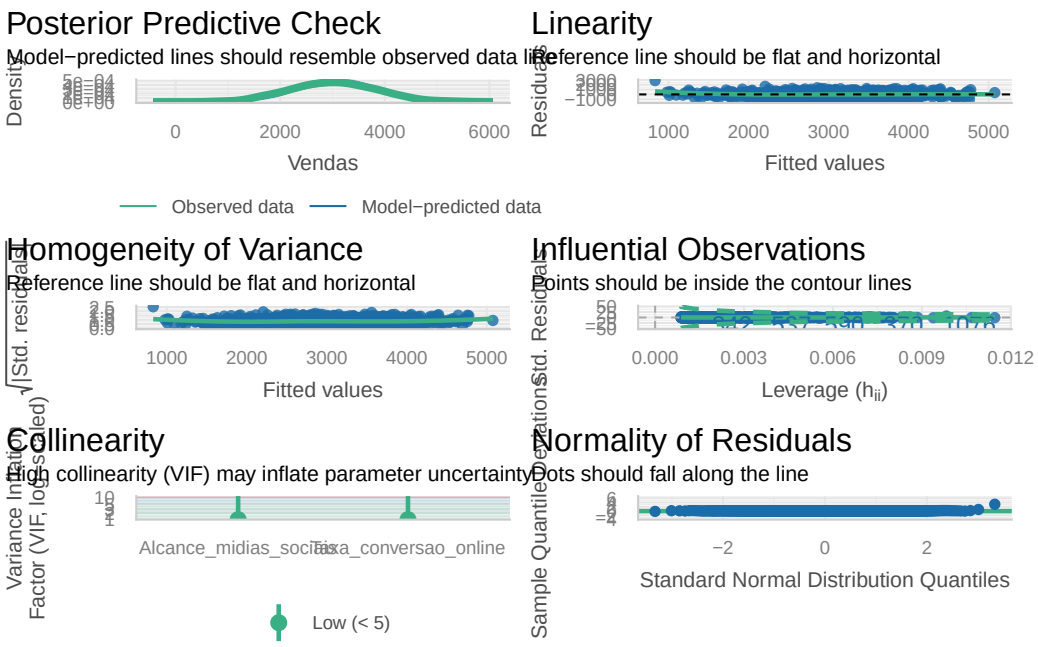
Existem alguns pontos numerados que se destacam por terem maior alavancagem (Leverage) ou resíduos maiores, como o ponto 1076. No entanto, todos os pontos estão bem dentro das linhas de contorno, o que significa que nenhum deles é considerado perigosamente influente.

6. Multicolinearidade

Mostra novamente os erros que já vimos, com um VIF **extremamente alto** de cerca de 15, dificultando a interpretação com o modelo.

5.1.2 Do modelo selecionado

```
check_model(modelo3, title_size = 10, base_size = 8)
```



Como o modelo 3 é simples, o gráfico de VIF foi omitido, justamente por não ter multicolinearidade nesse modelo, resolvendo esse problema em relação ao primeiro.

Nas influências de observações o único ponto que se encontrava fora da curva agora está contido nela, mostrando como o modelo 3 é mais estável, exatamente o que

queríamos.

Fora essas melhorias, o modelo 3 se mantém tão bom quanto o primeiro nas outras métricas, justificando novamente sua escolha.

6 Estimativa

```
estimativa <- data.frame(Alcance_midias_sociais = 1000, Taxa_conversao_onl

predict(modelo3, newdata = estimativa, interval = "confidence")
```

```
      fit      lwr     upr
1 3719.86 3655.861 3783.86
```

Considerando uma périodo de tempo onde foram atingidas 1000000 pessoas nas redes sociais, se espera que aproximadamente 3849 peças sejam vendidas nesse mês, com 95% de confiança se espera que o valor esteja entre 3808 e 3890.

6.1 Intervalo de confiança

limite	valor
inferior	3808
superior	3890

Esse intervalo de 3808 a 3890 expressa a incerteza em torno da média esperada do consumo para domicílios com as características especificadas. Ele reflete apenas o erro associado à estimativa da média, e não à previsão de um valor individual.

7 Previsão

```r
estimativa <- data.frame(Alcance_midias_sociais = 1000, Taxa_conversao_onl

predict(modelo3, newdata = estimativa, interval = "prediction")
```

```
       fit      lwr      upr
1 76873.48 48896.25 104850.7
```

Com 95% de confiança, espera-se que a quantidade de vendas de uma empresa com as características da seção 6 esteja entre 2960 e 4737. Esse intervalo de predição é mais amplo do que o intervalo de confiança para a média, pois leva em consideração a variabilidade individual das empresas, não apenas a incerteza da média.

8 Interpretação do modelo

Os resultados expostos mostram que todas as variáveis selecionadas para o modelo são estatísticamente significativas ao nível de 5%, resultando em um modelo com R^2 ajustado de 0.705, que indica que 70% da variabilidade no volume de vendar é explicado pelas variáveis no modelo. Esse valor é consideravelmente alto especialmente considerando quantas variáveis possíveis existiam no dataset, escolhemos um conjunto muito pequeno e significativo.

Análisando as variáveis incluídas: - `Alcance_midias_sociais` é a variável mais relevante para o modelo, mostrando como é importante esse meio de marketing para as lojas.

- `Taxa_conversao_online` tem uma influência bem menor, mas ainda significativa, possivelmente indicando que as pessoas compram mais online do que presencial.

Além disso, a análise dos pressupostos confirma que o modelo atende a todos os critério quase perfeitamente, sendo linear, independente de erros, atendendo a homocedasticidade e normalidade dos resíduos, com quase total ausência de multicolinearidade.

9 Conclusão

O relatório teve como objetivo analisar os fatores que influenciam no volume de vendas de peças de roupas. Após a limpeza de dados foram selecionadas variáveis explicativas com base em critérios estatísticos e teóricos.

O modelo de regressão linear múltipla selecionado (modelo 3) apresenta um bom ajuste (R^2 ajustado de 0.705) e atende aos pressupostos do modelo. A variável `Alcance_midias_sociais` apresenta uma influência significativa sobre o volume de vendas, com a variável `Taxa_conversao_online` tendo um impacto menor, mas considerável.

Por último, fizemos ainda estimativas e predições com base no modelo final, fornecendo interpretações pontuais e intervalares

10 Referências bibliográficas

1. *McElreath, Richard (2016). Statistical Rethinking: A Bayesian Course with Examples in R and Stan. [S.l.]: CRC Press. p. 189. ISBN 978-1-4822-5344-3. AIC provides a surprisingly simple estimate of the average out-of-sample deviance.* 2.https://www.datacamp.com/pt/tutorial/variance-inflation-factor