Reoferecimento de Estística
Case 9: Regressão Linear

Camille Peixoto Almeida
Isabela Belapetravicius

24 de Junho de 2023

1 Introdução

O presente estudo de caso tem como objetivo apresentar uma análise completa do tema **Regressão Linear** – desde uma abordagem teórica até aspectos mais práticos, incluindo a construção de intervalo de confiança, teste de hipóteses e tabela ANOVA.

Dentre as vantagens de se ajustar um modelo de regressão linear para um conjunto de dados, destaca-se a possibilidade de análise de certa tendência ou padrão, o que permite prever e estimar resultados futuros e desconhecidos.

Porém, a aplicabilidade de um modelo como esse implica a necessidade de certa linearidade dos dados, o que pode ser avaliado pelo R^2, denominado coeficiente de determinação, que ao se aproximar de 1 indica um maior comportamento linear.

Assim, a partir de um conjunto de dados, deseja-se ajustar um modelo linear e, utilizando as técnicas estatísticas vistas anteriormente no curso, avaliar seus coeficientes e determinar se há ou não correlação linear entre as duas variáveis escolhidas.

A base de dados de interesse contém informações sobre o investimento mensal, em milhares, de uma empresa em marketing para diversas plataformas e seu retorno em número de venda.

2 Desenvolvimento

2.1 Regressão Linear: Teoria

A princípio, define-se regressão linear como uma técnica estatística utilizada para modelar a relação entre uma variável aleatória dependente (eixo das ordenadas) e uma ou mais variáveis independentes (eixo das abscissas).

Aqui – a partir de conceitos de probabilidade – recorda-se que uma variável é dita aleatória quando a esta pode se associar uma função que retorna um número para cada resultado possível de um experimento aleatório. Essa variável pode ser discreta, assumindo um valor enumerável, ou contínua, assumindo quaisquer valores dentro de um intervalo.

Além disso, definem-se também experimentos determinísticos - aqueles em que o resultado é completamente previsível e não está sujeito a nenhum elemento aleatório – e experimentos probabilísticos – em que o resultado é incerto e depende de fatores aleatórios.

No contexto da regressão linear, a relação entre as variáveis é tratada como determinística, uma vez que se supõe que os parâmetros do modelo são fixos e não estão sujeitos a incerteza. Em contrapartida, os erros – isto é, a diferença entre os valores observados e os valores previstos pelo modelo – são aleatórios, caracterizando-os como elementos probabilísticos.

Por fim, são definidos os conceitos de parâmetro, estimador e estimativa. Um parâmetro é uma medida numérica que descreve uma característica da população, um estimador é uma função estatística aplicada a uma amostra, usada para estimar o valor de um parâmetro desconhecido, e uma estimativa é o valor numérico obtido ao aplicar um estimador a uma amostra específica.

Em um modelo de regressão linear simples, temos o parâmetro do coeficiente angular (representado por β_1) e o parâmetro do coeficiente linear (representado por β_0), ou seja:

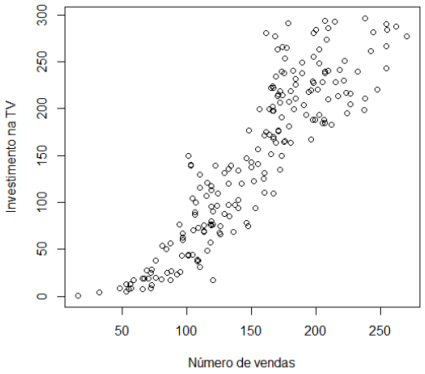
$$y = \beta_0 + \beta_1 x + e \tag{1}$$

Em relação aos estimadores, aqueles mais comumente utilizados na regressão linear são baseados no método dos mínimos quadrados, que busca minimizar a soma dos quadrados dos resíduos (denotados por e na expressão acima) entre os valores observados e os valores previstos pelo modelo. A partir dos estimadores, é possível obter estimativas para os parâmetros desconhecidos, como os coeficientes angular e linear do modelo de regressão linear.

Neste estudo, busca-se ajustar um modelo em que o número de vendas seja uma função linear do investimento em marketing pela TV.

Inicialmente, contrói-se o gráfico de dispersão do investimento em função do número de vendas, apresentado abaixo:

Figura 1: Dispersão do investimento na TV em função do número de vendas

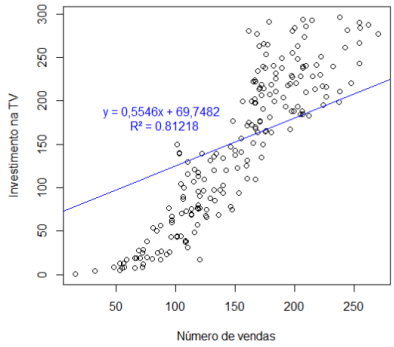


A partir do gráfico, nota-se que há certa relação linear entre as variáveis, à medida que o investimento na TV parece acompanhar o crescimento do número de vendas. Porém, para afirmar com propriedade, serão feitas análises mais aprofundadas posteriormente.

2.2 Regressão Linear: Prática

Agora, colocando os conceitos em prática, ajusta-se um modelo linear para o gráfico de dispersão do investimento na TV em função do número de vendas:

Figura 2: Modelo linear da dispersão do investimento na TV em função do número de vendas



Confirmando a suspeita anterior, a tendência apresentada pelo gráfico mostra uma relação crescente entre as variáveis. Além disso, os coeficientes angular e linear obtidos pelo modelo são, respectivamente, 0,5546 e 69,7482.

2.3 Teste de Hipóteses

Para determinar se há ou não correlação linear entre as variáveis analisadas, pode-se realizar um teste de hipóteses, em que:

- H_0: $\beta_0 = 0$ *e* $\beta_1 = 0$ *(Não há correlação)*
- H_1: $\beta_0 \neq 0$ *e* $\beta_1 \neq 0$ *(Há correlação)*

Os resultados obtidos estão apresentados a seguir:

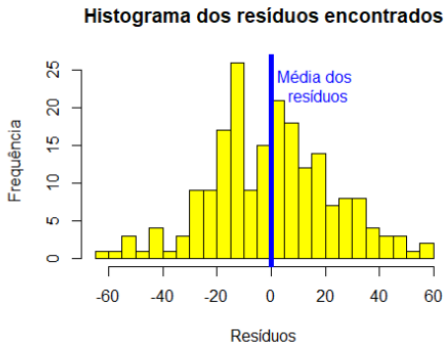
Coeficiente	**Estimativa**	**Erro**	**valor-t**	**valor-p**
Linear	69,7482	3,2255	21,62	<2e-16
Angular	0,5546	0,0190	29,34	<2e-16

Nota-se que o valor-t de cada coeficiente, calculado pela razão entre estimativa e erro, é muito maior que o valor-t referente a uma significância de 5%, definida como padrão no R, o que resulta em um valor-p extremamente pequeno (menor que 5%).

Assim, a hipótese nula é rejeitada, ou seja, pode-se afirmar que há uma correlação linear entre investimento na TV e número de vendas.

Além disso, o teste devolve também o valor de R^2, que mede a proporção da variabilidade da variável dependente – nesse caso, do investimento na TV – no modelo. O valor obtido para R^2 foi de 0,8122.

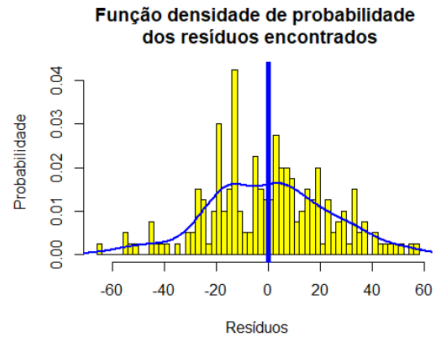
Em seguida, monta-se um histograma dos resíduos encontrados, exposto abaixo:



Sabe-se que os resíduos podem ser usados para avaliar a adequação do modelo estatístico aos dados, isto é, se os resíduos forem aleatórios, com média zero e variância constante, sugere-se que o modelo está capturando adequadamente a variabilidade dos dados.

Como a média dos resíduos encontrados é muito próxima de zero – cerca de -5×10^{-16} –, pode-se concluir que o modelo estatístico empregado é adequado para representar os dados.

Ainda, utilizando a função densidade de probabilidades, é possível visualizar a distribuição dos resíduos:



Além do teste de hipóteses para os coeficientes da reta de regressão linear, pode-se também realizar um teste para o coeficiente de correlação, caracterizado por uma medida estatística que avalia a força e a direção da relação linear entre duas variáveis quantitativas, variando entre -1 e 1.

Um coeficiente de correlação linear positivo ($0 < r < 1$) indica uma relação linear positiva, o que significa que à medida que os valores de uma variável aumentam, os valores correspondentes da outra variável também aumentam de maneira proporcional.

Por outro lado, um coeficiente de correlação linear negativo ($-1 < r < 0$) indica uma relação linear negativa. Logo, à medida que os valores de uma variável aumentam, os valores correspondentes da outra variável diminuem de maneira proporcional.

Para o caso de análise, o coeficiente de correlação vale $0,9012$.

Agora, as hipóteses são tais que:

- H_0*: $\rho = 0$ (Não há correlação)*
- H_1*: $\rho \neq 0$ (Há correlação)*

O teste de hipótese realizado com o auxílio do R resulta em um p-valor menor que $2,2 \times 10^{-16}$ e, portanto, muito menor que a significância de 5%. Assim, mais uma vez a hipótese nula é rejeitada, ou seja, pode-se afirmar que há uma correlação linear entre investimento na TV e número de vendas.

2.4 Intervalo de Confiança e ANOVA

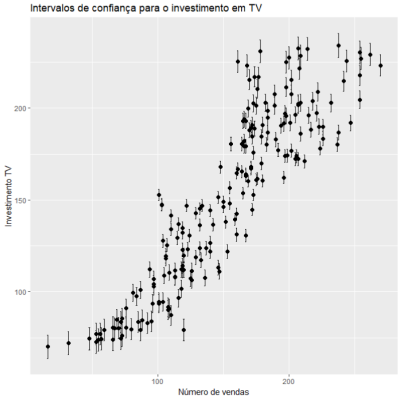
Na seção anterior, o valor obtido para o coeficiente de correlação representa uma estimativa – denotada por r – para o parâmetro ρ, que se refere ao valor populacional.

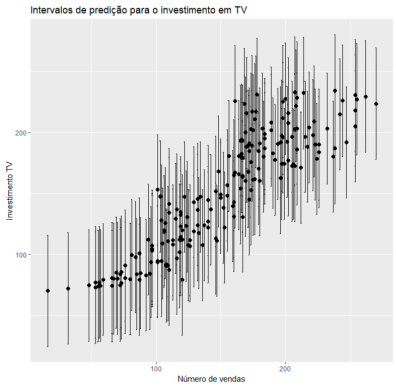
Assim, pode-se construir um intervalo de confiança centrado em r e que contenha ρ, considerando um nível de 5% de confiança.

O intervalo de confiança obtido com o auxílio do R foi tal que:

$$0,8714 \leq \rho \leq 0,9244 \tag{2}$$

Em seguida, são plotados os gráficos dos intervalos de confiança e predição para o investimento em TV:





A partir dos gráficos, observa-se que o intervalo de confiança é muito menor quando comparado ao intervalo de predição. Isso porque o intervalo de confiança contém o parâmetro de interesse – neste caso, o coeficiente correlação –, que é fixo e não está sujeito a incerteza, enquanto o intervalo de predição considera os erros do modelo, que são aleatórios e seguem uma distribuição de probabilidades, conforme exposto na introdução.

Finalmente, monta-se a tabela de Análise de Variâncias (ANOVA):

Fonte	**SQ**	**GL**	**QM**	**Fcalc**	**Fcrit**
Regressão	451244	1	451244	856,2	2,734
Residual	104355	198	527		
Total	555599	199			

Aqui, lembra-se que

$$QM = \frac{SQ}{GL} \tag{3}$$

$$F_{calc} = \frac{QM_{Regressao}}{QM_{Residual}} \tag{4}$$

$$F_{crit} = F_{GL_{Regressao};GL_{Residual};5\%} \tag{5}$$

Considerando novamente as hipóteses:

- H_0*:* $\beta_0 = 0$ *e* $\beta_1 = 0$ ***(Não há correlação)***
- H_1*:* $\beta_0 \neq 0$ *e* $\beta_1 \neq 0$ ***(Há correlação)***

Conclui-se que a hipótese nula é rejeitada, uma vez que $F_{calc} >> F_{crit}$. Portanto, mais uma vez pode-se afirmar que há uma correlação linear entre investimento na TV e número de vendas.

Além disso, o coeficiente de determinação R^2 pode ser calculado através da ANOVA, pela relação:

$$R^2 = \frac{SQ_{Regressao}}{SQ_{Total}} = \frac{451244}{555599} = 0,8122 \tag{6}$$

Código de R

A seguir, expõe-se o trecho de código utilizado para realizar este estudo de caso:

```
df_vendas <- readRDS("H:/Meu
Drive/USP/semestres_passados/1°Quadri2023/reof_estat/Estudo de Caso 9 - Regressão
Linear-20230614/Case9/vendas.rds")

# gráfico de dispersão
plot(df_vendas$vendas, df_vendas$tv, ylab = "Investimento na TV", xlab = "Número de
vendas")

# regressão linear
help("lm")
modelo1 <-lm(df_vendas$vendas~df_vendas$tv, data = df_vendas)

# coeficientes
abline(modelo1, col = "blue") + text(x = 90, y = 180, col = "blue","y = 0,5546x + 69,7482 \n R²
= 0.81218")

# teste de hipóteses
summary(modelo1)

# histograma
df_residuals <- modelo1[["residuals"]]
media_residuos <- mean(df_residuals)

hist(df_residuals,
    breaks = 35,
    freq = T,
    col = "yellow",
    ylab = "Frequência",
    xlab = "Resíduos",
    main = "Histograma dos resíduos encontrados",
    text(x = 30, y = 20, "média"))
abline(v = media_residuos, lwd = 5, col = "blue")
text(x = media_residuos + 17, y = 23, "Média dos \n resíduos", col = "blue")

# função de probabilidade
hist(df_residuals,
    breaks = 50,
    freq = F,
    col = "yellow",
    ylab = "Probabilidade",
    xlab = "Resíduos",
    main = "Função densidade de probabilidade \n dos resíduos encontrados",
    text(x = 30, y = 20, "média"))
densidade <- density(df_residuals)
```

```
lines(densidade, col = "blue" , lwd = 2)
abline(v = media_residuos, lwd = 5, col = "blue")
text(x = media_residuos + 10, y = 23, "Média dos\n resíduos", col = "blue")

# correlação
help(cor)
cor(df_vendas$vendas, df_vendas$tv)

# intervalos de predição e confiança
tv = data.frame(tv = sort(df_vendas$tv))

estimados_IC <- predict(modelo1, tv, interval = "confidence")
previstos_IP <- predict(modelo1, tv, interval = "prediction")

df_estimados_IC <- read_excel("H:/Meu
Drive/USP/semestres_passados/1°Quadri2023/reof_estat/Estudo de Caso 9 - Regressão
Linear-20230614/estimados_IC.xlsx")

df_previstos_IP <- read_excel("H:/Meu
Drive/USP/semestres_passados/1°Quadri2023/reof_estat/Estudo de Caso 9 - Regressão
Linear-20230614/previstos_IP.xlsx")

ggplot(df_estimados_IC, aes(x = vendas, y = fit)) +
  geom_point(size = 2)+
  labs(title ="Intervalos de confiança para o investimento em TV",
       x = "Número de vendas",
       y = "Investimento TV")+
  geom_errorbar(aes(ymin = lwr, ymax = upr))

ggplot(df_previstos_IP, aes(x = vendas, y = fit)) +
  geom_point(size = 2)+
  labs(title ="Intervalos de predição para o investimento em TV",
       x = "Número de vendas",
       y = "Investimento TV")+
  geom_errorbar(aes(ymin = lwr, ymax = upr))

# ANOVA
aov <- aov(modelo1)
summary (aov)
```