Introdução à datascience com R

Cleuton Sampaio

Lição 7: Regressão

Regressão

Esta aula é sobre regressão. O que é regressão? Segundo a Wikipedia:

> *Em estatística, regressão é uma técnica que permite explorar e inferir a relação de uma variável dependente (variável de resposta) com variáveis independentes específicas (variáveis explicatórias). A análise da regressão pode ser usada como um método descritivo da análise de dados (por exemplo, o ajustamento de curvas) sem serem necessárias quaisquer suposições acerca dos processos que permitiram gerar os dados. Regressão designa também uma equação matemática que descreva a relação entre duas ou mais variáveis.*
>
> *O método de estimação mais amplamente utilizado é o método dos mínimos quadrados ordinários.*
>
> *Os principais problemas que devem ser enfrentados em uma regressão são: multicolinearidade, heteroscedasticidade, autocorrelação, endogeneidade e atipicidade.*

Eu não descreveria melhor.

Antes de continuar, volte à sessão 1, reveja o vídeo, os slides e o PDF explicativo. Abra o script da aula 1 no RStudio, rode e procure relembrá-lo. Onde achar?

- Vídeo da sessão 1: https://youtu.be/E5MXqejCmaw
- Slides da sessão 1: https://github.com/cleuton/datascience/blob/master/R-course/lesson1/aula1-intro.pdf
- PDF da sessão 1: https://github.com/cleuton/datascience/blob/master/R-course/lesson1/regressao-linear-r.pdf
- Script R da sessão 1: https://github.com/cleuton/datascience/blob/master/R-course/lesson1/lerOds.R

Sério! Se não fizer isso, então pode parar o curso aqui mesmo.

Depois de ler a sessão 1 e rodar o exemplo novamente

Ótimo! Agora, você entende o suficiente de R para poder analisar melhor aquele script: "lerOds.R":

```
library('readODS')
library('tidyverse')
data <- read_ods('mod-preditivo.ods', sheet=2,col_names =
TRUE,range='a1:b30',col_types=NA)
print(data)
df <- type_convert(data,trim_ws=TRUE,col_types =
cols(Pesos=col_integer(),Alturas=col_double()),locale = locale(decimal_mark =
","))
str(df)
y <- df$Pesos
x <- df$Alturas
model <- lm(y ~ x)
summary(model)
df2 <- data.frame(x=c(1.40,1.90))
pesos2 <- predict(model,newdata = df2)
head(pesos2)
plot(x, y)
lines(df2$x,pesos2,col="red")
```

Ok, talvez você esteja em dúvidas sobre o pacote "tidyverse" e a função "type_convert()", que pertence a ele. Ela serve para converter os valores do formato Brasileiro para o R. Você já sabe como fazer isso dentro da função "read_csv()", não? Lembra da aula sobre Datasets externos? Ok. Outra coisa foi o "read_ods()", que pertence ao pacote "readODS", que eu utilizei para ler diretamente da planilha LibreOffice. Não há muito com o que se preocupar, pois é autoexplicativo. Não faz parte do objetivo deste curso, pois aqui, vamos ler apenas arquivos CSV.

O resto, incluindo as funções "lm()" e "summary()" você já conhece, certo?

Avaliando modelos de regressão

No R usamos a função "lm()" para criar modelos de regressão linear, baseados no método dos “Mínimos quadrados”.

Na aula 1, vimos como avaliamos nosso modelo pelo coeficiente de determinação ou R-quadrado. Em R, podemos saber qual é o R2 lendo o resultado da função “summary()”:

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04096 -1.01008  0.07419  1.00118  2.62196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -93.417      4.627  -20.19   <2e-16 ***
x             95.786      2.796   34.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.418 on 27 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9775,  Adjusted R-squared:  0.9767
F-statistic:  1174 on 1 and 27 DF,  p-value: < 2.2e-16
```

(Se o resultado da “summary()” não aparecer na console, coloque a chamada dentro de um “print()”: print(summary(model)))

Resíduos

A primeira parte está relacionada com os resíduos, que são as diferenças entre os valores observados e os valores estimados pelo nosso modelo. Em um bom modelo, a distribuição dos resíduos deve se aproximar da distribuição normal. Esta é uma das condições de avaliação de uma regressão linear.

A função “summary()” nos mostra isso:

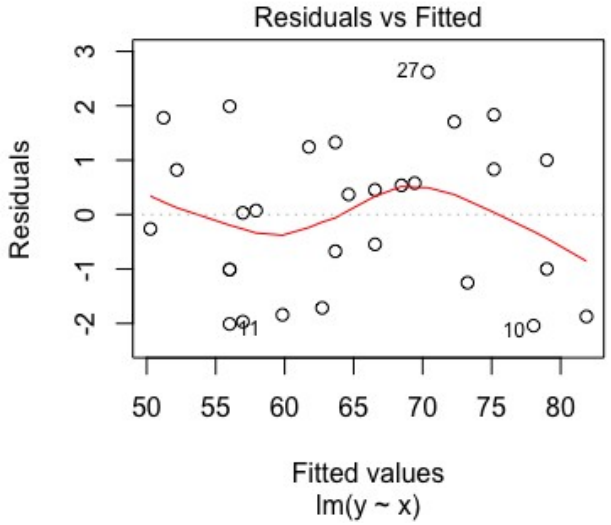
```
Residuals:
     Min       1Q   Median       3Q      Max
-2.04096 -1.01008  0.07419  1.00118  2.62196
```

Temos 5 pontos de resíduos: Mínimo, primeiro quartil, mediana, terceiro quartil e máximo. A distribuição dos resíduos em torno da média deve ser normal (sino). Podemos ver isso nos nossos valores, pois a mediana está bem próxima de zero e os outros valores

parecem ser bem simétricos. Podemos plotar a distribuição para ver isso. Rode este comando na console (após ter rodado o script “lerOds.R”):

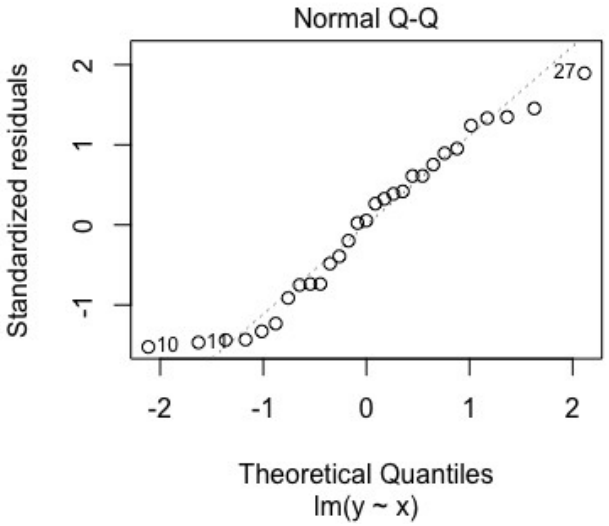
plot(model)

O primeiro gráfico mostra um possível problema:



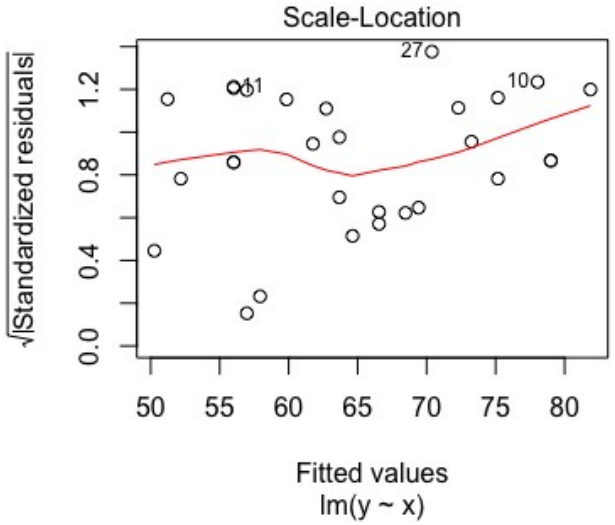
Ele deveria mostrar uma linha reta… Mas isto não significa que a relação seja diferente de linear, pois há outros indícios do contrário. Talvez, o tamanho da amostra esteja pequeno. Se você vir uma parábola, então, claramente o modelo linear não se aplicaria.

O segundo gráfico mostra “Normal Q-Q” mostra isso:



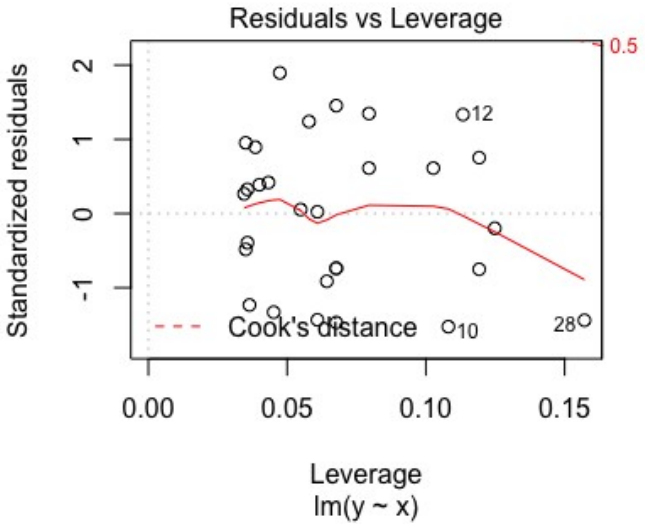
Ele deveria seguir a linha diagonal, mas mostra um pequeno desvio. Isso significa que apresenta uma pequena assimetria, o que também pode ser devido ao tamanho reduzido da amostra.

Já o terceiro gráfico “Scale-Location” mostra também uma situação pouco desejável:



Ele deveria mostrar uma linha aproximadamente reta, com os resíduos distribuídos aleatóriamente, sem tendências. Bem, a linha não está totalmente reta, mas também não é totalmente curva, e os resíduos aparentam estar aleatoriamente distribuídos em torno dela, mas há alguns “outliers” à esquerda, que dão um formato trapezoidal ao gráfico, e isto pode denotar um problema chamado de Heterocedasticidade (veremos adiante).

Finalmente, o quarto gráfico “Residuals vs Leverage” mostra o quanto os “outliers” podem ter influenciado o resultado da regressão:



Devemos procurar pontos muito afastados no canto superior ou inferior, pois estes poderiam alterar o resultado. Geralmente, são mostradas duas linhas tracejadas, chamadas de “Distância de Cook” (Cook’s distance). Se um ponto estiver fora dos limites destas linhas, então é um “outlier” que está alterando ou influenciando a regressão, devendo ser retirado da amostra. Nesta imagem, quase não podemos ver as linhas da distância de Cook, o que é um bom sinal.

Temos alguns probleminhas com os resíduos, mas eles parecem estar distribuídos de forma normal, o que confirmaria o modelo. Vimos que nossa amostra está ligeiramente assimétrica e isto pode causar estes problemas.

A avaliação mais importante é se os resíduos parecem estar simétricos em torno da média (zero).

Coeficientes

O resultado da função “summary()” mostra isso:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -93.417      4.627  -20.19   <2e-16 ***
x             95.786      2.796   34.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Temos dois coeficientes calculados: Intercept (coeficiente linear) e “x” (o multiplicador de “x” ou coeficiente angular, ou “slope”).

A primeira informação é o próprio valor dos coeficientes, na coluna “estimate”. Isto nos dá um modelo:

y = 95,786x – 93,417

A outra informação é o erro padrão (“std. Error”), que é o quanto a estimativa do coeficiente varia. Queremos um número pequeno. Nosso modelo estima o peso com base na altura, logo, uma variação de 1 metro equivale a 95,786 kilos de peso, o que parece muito, mas lembre-se que estamos medindo em metros e que raramente uma pessoa mede mais de 2 metros. Se fosse em centímetros, poderíamos ter coeficientes menores.

Podemos dizer que o peso de uma pessoa pode variar 2,786 kg.

O t-value é o quanto a nossa estimativa de coeficiente está longe da média zero, em desvios padrões. Queremos que esteja bem longe, para rejeitar a hipótese nula, de que o coeficiente seria zero. Se o valor for muito próximo de zero, significa que não existiria relacionamento entre a variável e a previsão.

O p-value é a probabilidade do teste da hipótese de que o coeficiente seria zero. Neste caso, não haveria relacionamento entre as variáveis. Se for uma regressão múltipla, com várias variáveis independentes, significaria que aquela seria irrelevante para a previsão, e poderia ser retirada.

No nosso caso, os p-values são muito pequenos e, certamente, menores que 0,05 (lembra do teste de hipótese a 95%? É isso mesmo!

Note que há 3 asteriscos ao lado dos p-values, indicando que os coeficientes são significativos para a regressão. Há até uma legenda para isso:

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Um ponto ou um branco indicam que o coeficiente não é significativo para a predição.

Erro padrão dos resíduos

Temos um “Residual Standard Error”:

```
Residual standard error: 1.418 on 27 degrees of freedom
```

Isto significa o quanto a resposta da regressão (o peso) pode se desviar da reta de regressão, que seria 1, 42 kg. Como a média de peso da amostra é de 64,86 kg (é só dar um “print(summary(df))”) isso representa uma variação de 2,18%.

Coeficiente de determinação

Temos os valores:

```
Multiple R-squared:  0.9775,  Adjusted R-squared:  0.9767
```

O valor “Multiple R-squared” é o valor do R-quadrado: 0,98, significando que 98% da variância do peso é explicada pela altura (nosso modelo), o que muito bom.

O “Adjusted R-squared” é o valor do R-quadrado pode aumentar artificialmente em regressões múltiplas,quando temos mais de uma variável independente (por exemplo, além da altura, usamos também a idade). Neste caso, podemos usar o R-quadrado ajustado (“Adjusted R-squared”), cuja fórmula é:

$$\bar{R}^2 = 1 - \frac{n-1}{n-(k+1)}(1-R^2)$$

Onde “k” é a quantidade de estimadores (variáveis independentes) e o 1 somado é a constante (coeficiente linear).

Significância da regressão

É um teste de hipótese para a hipótese nula de que todos os coeficientes seriam zero:

```
F-statistic:  1174 on 1 and 27 DF,  p-value: < 2.2e-16
```

O ideal é que o valor “F-statistic” seja bem maior que 1, o que, no nosso caso é. E queremos que o p-value seja próximo de zero, o que também é. Logo, rejeitamos a hipótese nula de que todos os coeficientes seriam zero.

Regressão múltipla

Está fora do escopo deste curso, mas seria algo como usar altura e idade para tentar calcular o peso. Em uma regressão múltipla, temos esta fórmula:

$y = ß_0 + x_1ß_1 + x_2ß_2 + \ldots + x_nß_n$

Cada valor Beta (ß) é um coeficiente a ser estimado pela função de regressão. O resultado desta equação não é uma reta, mas um plano.

Todos os cuidados que tomamos com a avaliação da regressão simples tornam-se críticos na regressão múltipla.

Regressão não linear

É uma análise de regressão cujo modelo é uma função não linear dos estimadores. Existem vários modelos de regressões que podem ser aplicados.

Se você observar evidências de que a relação não seja linear (como já mostramos) talvez seja melhor utilizar outro método de Machine Learning, como Árvores de Decisão, por exemplo.

Este assunto foge do escopo deste curso.

Restrições de regressões lineares

Para serem válidas, as regressões lineares precisam seguir quatro restrições importantes:

1. A média da variável dependente (E(Y)) é obtida através de uma função linear entre os valores das variáveis independentes;
2. Os erros de previsão (desvios não explicados ou ($y_i - \hat{y}_i$) são independentes, ou seja, o erro de uma observação não influencia o erro de outra observação;
3. Os erros em cada observação possuem distribuição normal;
4. Os erros em cada observação possuem variâncias iguais (σ^2).

Problemas que podem ocorrer

Multicolinearidade

Ocorre em regressões multivariadas, quando duas ou mais variáveis independentes apresentam correlação entre si. Há alguma discussão sobre os efeitos da multicolinearidade, já que pode ser muito difícil reduzi-la. Porém, aceita-se que multicolinearidade severa gerar erros padrões elevados e afetar a estabilidade dos coeficientes.

Heterocedasticidade

Uma das condições necessárias para a validade das inferências de uma regressão é que o termo do erro aleatório, ε, tenha uma variância constante para todos os níveis de variáveis independentes.

Quando esta condição é satisfeita, o modelo é dito Homocedástico. Quando são observadas variações desiguais para diferentes conjuntos de variáveis independentes, o modelo é dito Heterocedástico.

Autocorrelação dos resíduos

Significa que o erro de hoje está influenciando o de amanhã. É uma situação em que os resíduos estão correlacionados, e ocorre frequentemente em séries de dados temporais.