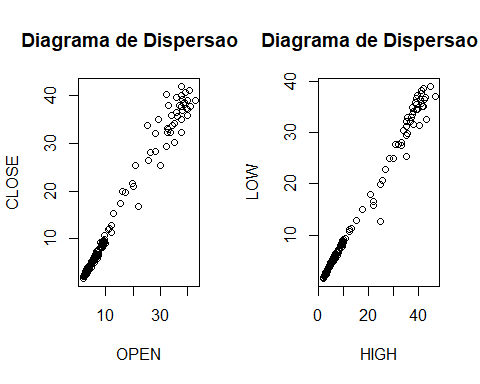
Atividade de Calculo - Modelagem no RStudio

Rita de Cássia

2023-11-19

## Obtendo os dados

base <- read.csv("C:/Users/ritad/Desktop/Faculdade/Aprendendo\_Calculos/Aulas de R/Exercicios/base-r-wege-full.csv")  
  
par(mfrow = c(1, 2))  
plot(base$Open, base$Close, main = "Diagrama de Dispersao A", xlab = "OPEN", ylab = "CLOSE")  
plot(base$High, base$Low, main = "Diagrama de Dispersao B", xlab = "HIGH", ylab = "LOW")



#### Breve Análise:

É possível observar que esses dados têm uma forte correlação positiva entre eles (proporção direta entre os dados), pois não há muita dispersão, apenas no gráfico B que ocorre uma porcentagem maior de dispersão. Existem alguns outliers, mas são poucos

### Aproximando uma reta

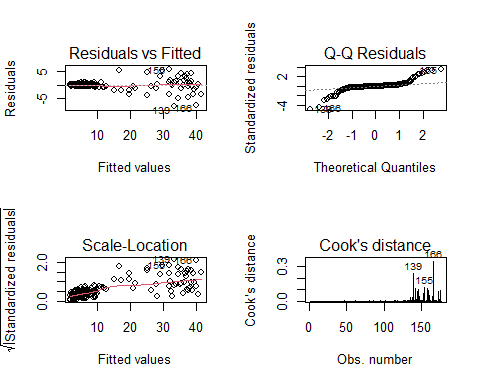
#Criando os modelos   
modeloWegeO <- lm(base$Open ~ base$Close)  
summary(modeloWegeO)

##   
## Call:  
## lm(formula = base$Open ~ base$Close)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8459 -0.2938 -0.0235 0.2800 5.6908   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.002186 0.175520 -0.012 0.99   
## base$Close 0.987321 0.009781 100.947 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.666 on 175 degrees of freedom  
## Multiple R-squared: 0.9831, Adjusted R-squared: 0.983   
## F-statistic: 1.019e+04 on 1 and 175 DF, p-value: < 2.2e-16

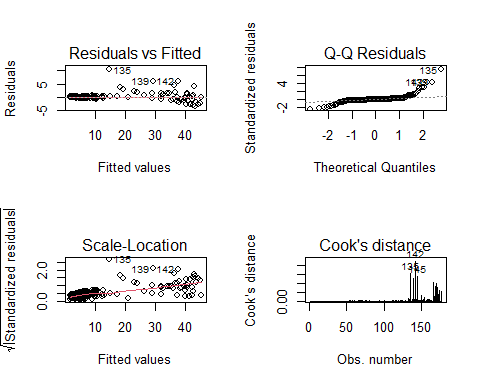
#Criando os modelos   
modeloWegeH <- lm(base$High ~ base$Low)  
summary(modeloWegeH)

##   
## Call:  
## lm(formula = base$High ~ base$Low)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4042 -0.3443 -0.1535 0.1103 10.3444   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.001659 0.147283 -0.011 0.991   
## base$Low 1.159481 0.008875 130.649 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.399 on 175 degrees of freedom  
## Multiple R-squared: 0.9899, Adjusted R-squared: 0.9898   
## F-statistic: 1.707e+04 on 1 and 175 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(modeloWegeO, which = 1:4)



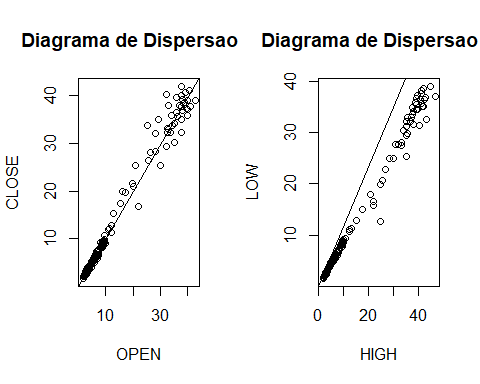
plot(modeloWegeH, which = 1:4)



#### Breve Análise:

A partir da criação desse modelo e do summary dos dois gráficos, podemos confirmar a possibilidade de correlação positiva, nesse modelo o “Multiple R-squared” está acima de 0.97, ou seja 97%, e quanto mais próximo de 1 ele estiver menor é a probabilidade de erro ao confiar nesse modelo. Com esse modelo também podemos ver se existem outliers e como foi disse anteriormente, eles são poucos

par(mfrow = c(1, 2))  
plot(base$Open, base$Close, main = "Diagrama de Dispersao A", xlab = "OPEN", ylab = "CLOSE")  
abline(modeloWegeO)  
  
plot(base$High, base$Low, main = "Diagrama de Dispersao B", xlab = "HIGH", ylab = "LOW")  
abline(modeloWegeH)

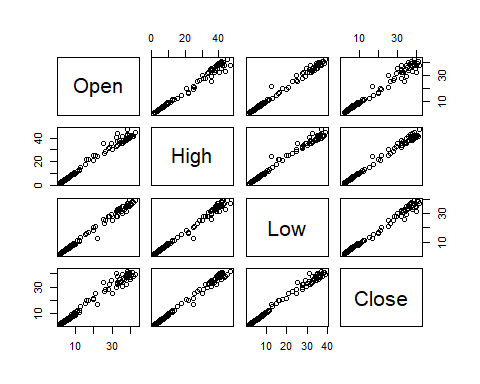


#### Breve Análise:

O Diagrama A segue perfeitamente a linha de modelo encontrada, mas o B se dispersa principalmente ao final o que pode ser a indicação de que esse modelo não é tão adequado

## Realizando uma regressão multipla linear

base\_completa <- base[,c("Open", "High", "Low", "Close")]  
  
pairs(base\_completa)



#### Breve Análise:

Agora com a regressão multipla linear é possível observar que os dados tem uma forte correlação positiva entre eles, pois todos formam uma relação linear ou quase linear

library(GGally)

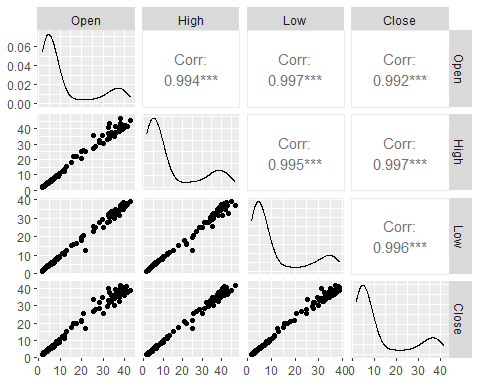
## Warning: package 'GGally' was built under R version 4.3.2

## Carregando pacotes exigidos: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.2

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

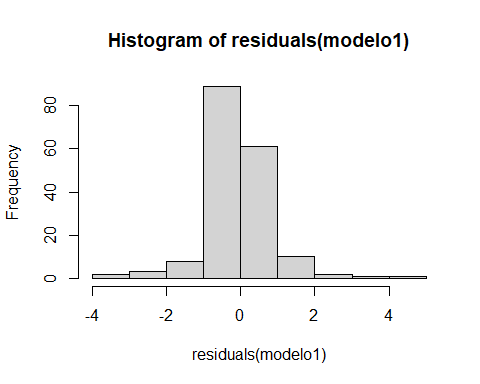
ggpairs(base\_completa)



summary(base\_completa)

## Open High Low Close   
## Min. : 1.701 Min. : 1.782 Min. : 1.531 Min. : 1.723   
## 1st Qu.: 3.339 1st Qu.: 3.567 1st Qu.: 3.244 1st Qu.: 3.451   
## Median : 6.158 Median : 6.665 Median : 5.881 Median : 6.169   
## Mean :12.415 Mean :13.468 Mean :11.617 Mean :12.576   
## 3rd Qu.:16.200 3rd Qu.:20.915 3rd Qu.:15.065 3rd Qu.:17.330   
## Max. :42.705 Max. :46.930 Max. :39.005 Max. :41.895

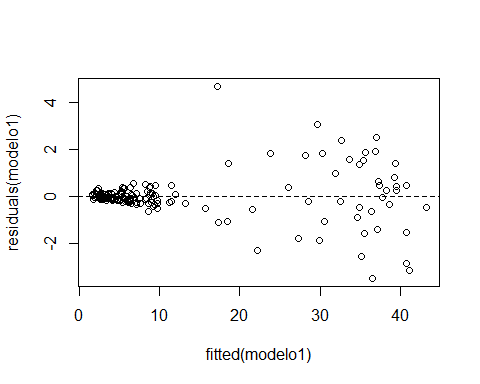
modelo1 <- lm(Open ~ High + Low + Close, data = base\_completa)  
  
hist(residuals(modelo1))



#### Breve Análise:

É possível ver que há uma distribuição normal desses dados e para confirmar faremos um gráfico de valor ajustado versus resíduo

plot(fitted(modelo1), residuals(modelo1))  
abline(h = 0, lty = 2)



#### Breve Análise:

Com esse gráfico podemos confirmar que esses resíduos estão bem distribuídos pela linha, o que ajuda na confirmação de confiança do modelo

summary(modelo1)

##   
## Call:  
## lm(formula = Open ~ High + Low + Close, data = base\_completa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5128 -0.1899 -0.0533 0.1261 4.7093   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.004516 0.095583 0.047 0.962   
## High 0.524329 0.064438 8.137 7.61e-14 \*\*\*  
## Low 1.002894 0.066194 15.151 < 2e-16 \*\*\*  
## Close -0.501096 0.076380 -6.561 5.98e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9062 on 173 degrees of freedom  
## Multiple R-squared: 0.9951, Adjusted R-squared: 0.995   
## F-statistic: 1.162e+04 on 3 and 173 DF, p-value: < 2.2e-16

sigma(modelo1)/mean(base\_completa$Open)

## [1] 0.072993

#### Breve Análise:

* O summary apresenta que não há uma grande variabilidade desses dados a partir do (>|t|). O modelo se dá útil.
* O Multiple R-squared em 0.9951 siginifica que 99% da variância em Open pode ser explicada pelos preditores do modelo.
* O Residual standard error representa que os valores apresentados caem em média 0.9062 unidades da linha de regressão, ou seja, muito pouco, sendo uma medida de erro de previsõa baixa
* Esse modelo apresenta cerca de 7,29 % de erro de acordo com o sigma

## Conclusão

* A partir dessa base dados, é possível observar que o modelo que mais se adequa é o modelo de regressão linear, a partir da variável de coluna Open dentro da base de dados. De acordo com a divisão feita no início dos diagramas de dispersão, é notável uma forte correlação positiva entre as variáveis chamadas “Open”, “Close” e “High”,“Low” e com os modelos de regressão linear simples foi visto um alto valor no R-Squared o que mostra uma boa capacidade do modelo de explicar a variabilidade dos dados, os summaries, o cook distance e o hist facilitaram a observação de poucos outliers e finalmente é possível confirmar as suposições de que os modelos têm um bom desempenho tanto na explicação quanto na previsão dos dados