Lista 5 - MAE 0399 - Análise de Dados e Simulação

Guilherme Ventura (11340293), Milton Leal (8973974), Richard Sousa (11810898)

17/07/2021

Questão 2

Item a) R:

A fórmula geral para a probabilidade do modelo simples de regressão logística com k variáveis preditoras e com base nos parâmetros estimados é dada por:

$$\hat{p}(x) = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)}$$

No caso do exercício, temos $\beta_0 = -6$, $\beta_1 = 0.05$ e $\beta_2 = 1$. Além disso, $x_1 = 40$ e $x_2 = 3.5$.

Portanto,

$$\hat{p}(y = 1 | x_1 = 40, x_2 = 3.5) = \frac{exp(-6 + 0.05 * 40 + 1 * 3.5)}{1 + exp(-6 + 0.05 * 40 + 1 * 3.5)} = 0.3775$$

Item b) R:

Queremos que $\hat{p}(y = 1 | x_1 = 40, x_2 = 3.5) = 0.5$, considerando que a nota do estudante seja igual a 3.5. Então, temos:

$$0.5 = \frac{exp(-6 + 0.05 * x_1 + 1 * 3.5)}{1 + exp(-6 + 0.05 * x1 + 1 * 3.5)} =$$

$$0.5 + \frac{exp(-6 + 0.05 * x_1 + 1 * 3.5)}{2} = exp(-6 + 0.05 * x_1 + 1 * 3.5) =$$

$$1 = exp(-2.5 + 0.05 * x_1) =$$

$$log(1) = log(exp(-2.5 + 0.05 * x_1)) =$$

$$0 = -2.5 + 0.05 * x_1 =$$

$$x_1 = 50$$

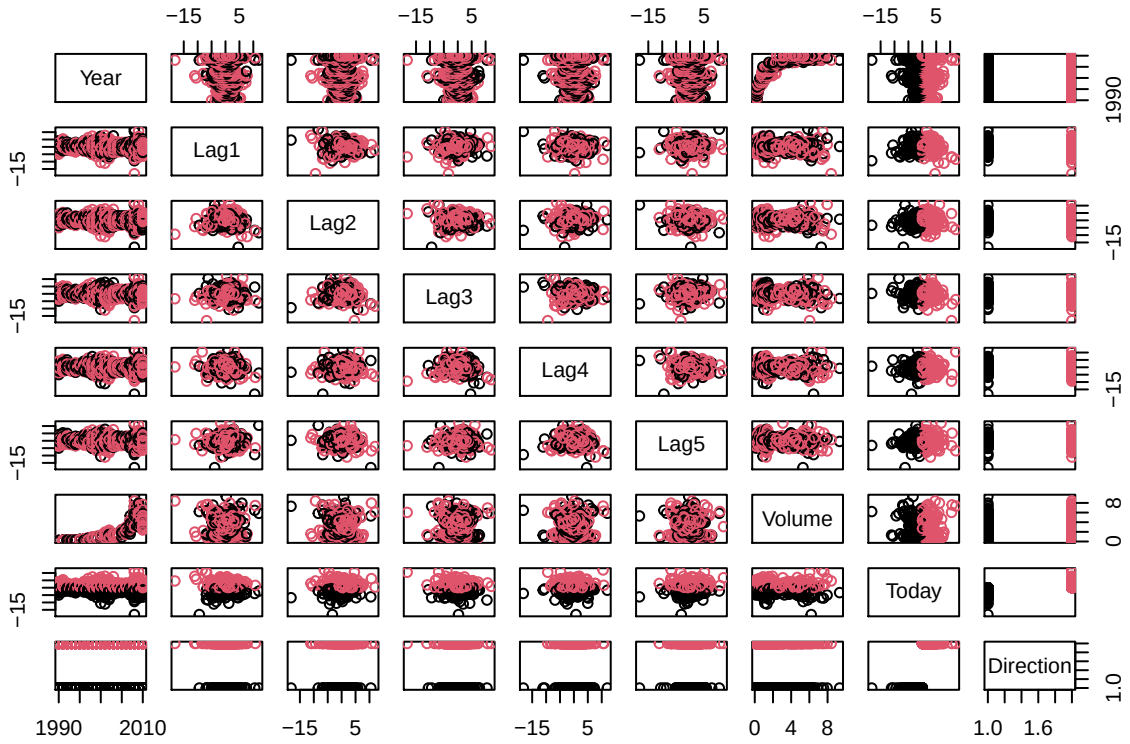
Portanto, para ter probabilidade de 50% de obter nota A na disciplina, o estudante deve estudar 50h.

Questão 4

```
library(ISLR)
attach(Weekly)
```

a)

```
plot(Weekly, col=Weekly$Direction)
```



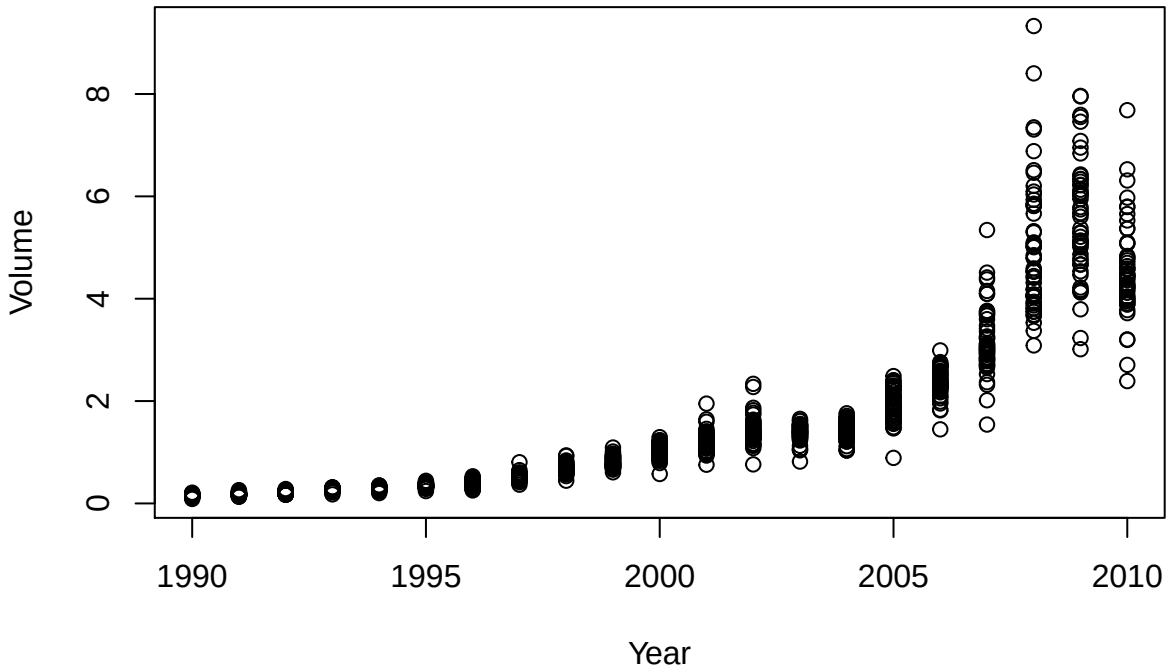
O gráfico de scatterplot acima nos mostra uma visão geral entre as variáveis levando em consideração se o mercado subiu (cor vermelha) ou desceu (cor preta) em determinada semana. Observamos que os gráficos não conseguem nos dizer muita coisa, não existe uma relação clara de correlação entre as variáveis.

```r
cor(Weekly[,-9])
```

```
##               Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```

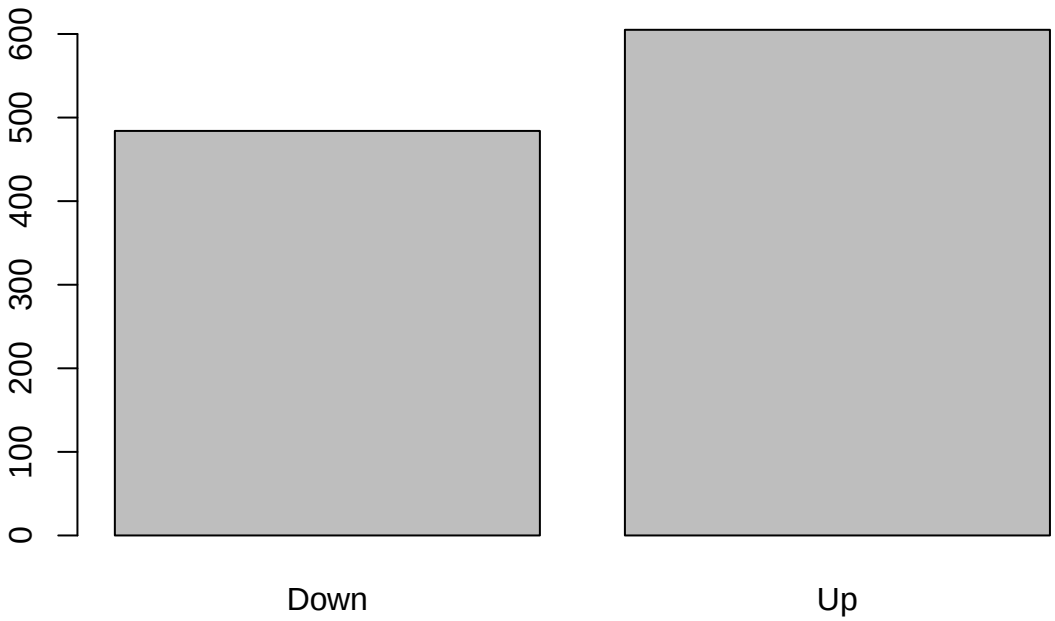
A partir da matriz de correlação acima, podemos confirmar a não existência de correlação entre a maioria das variáveis a partir da matriz de correlação. Observamos que apenas as variáveis YEAR e VOLUME possuem correlação significativa.

```
plot(Year, Volume)
```



É possível perceber que o volume de negociações tem crescido ao longo dos anos, mas que também houve aumento na dispersão dos valores dentro de um determinado ano ao longo do tempo.

```
barplot(table(Direction))
```



Podemos observar que a variável resposta Direction apresenta mais quantidades de Up do que de Down, mostrando que o mercado financeiro subiu mais do que desceu nas semanas em que foram coletados os dados.

b)

```
ajuste_reglog = glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=Weekly, family = binomial)
summary(ajuste_reglog)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

Analisando os resultados, percebemos que apenas o intercepto e o preditor Lag2 são estatisticamente significantes, pois apresentam p-valor abaixo de 10% para um nível de significância de 10%.

Neste teste, as hipóteses são usadas para testar se os coeficientes associados ao intercepto e as covariáveis são zero ou não.

Então, temos $H_0 : \beta_i = 0$ e $H_1 : \beta_i \neq 0$ com $i \in \{0, 1, 2, 3, 4, 5, 6\}$.

Neste caso, rejeitamos todas as hipóteses nulas com exceção de β_0 e β_2.

c)

```
glm.probs=predict(ajuste_reglog, type="response")
glm.pred=ifelse(glm.probs>0.5, "Up", "Down")

#Abaixo, temos a matriz de classificação
table(glm.pred, Direction)
```

```
##         Direction
## glm.pred Down  Up
##     Down   54  48
##     Up    430 557
```

```
#Total de Up
total_up = 48+557
total_up
```

```
## [1] 605
```

```
#Total Down
total_down = 54+430
total_down
```

```
## [1] 484
```

```
#Sensibilidade (considerando UP como a classe positiva)

sensibilidade = 557/total_up
#Abaixo, temos a taxa do modelo classificar como Up quando ele realmente é Up.
sensibilidade
```

```
## [1] 0.9206612
```

```
#Especificidade

especificidade = 54/total_down
#Abaixo, temos a taxa do modelo classificar como Down quando ele realmente é Down.
especificidade
```

```
## [1] 0.1115702
```

```
#Total de predições corretas

total_correto = (54+557)/(total_down + total_up)
#Abaixo, temos a fração total de predições corretas
total_correto
```

```
## [1] 0.5610652
```

d)

```
treino = Year < 2009
ajuste_reglog2 = glm(Direction~Lag2, data= Weekly, family = binomial, subset = treino)
summary(ajuste_reglog2)
```

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = Weekly,
##     subset = treino)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.536  -1.264   1.021   1.091   1.368
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.20326    0.06428   3.162  0.00157 **
## Lag2         0.05810    0.02870   2.024  0.04298 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1354.7  on 984  degrees of freedom
## Residual deviance: 1350.5  on 983  degrees of freedom
## AIC: 1354.5
##
## Number of Fisher Scoring iterations: 4
```

Podemos observar acima que tanto o intercepto quanto a variável preditora Lag2 são estatisticamente sginificantes a um nível de significância de 10%.

Vamos agora considerar como amostra de teste os anos seguintes (2009 e 2010) do conjunto de dados Weekly.

```
glm.probs2=predict(ajuste_reglog2, newdata = Weekly[!treino,],type="response")
glm.pred2=ifelse(glm.probs2>0.5, "Up", "Down")
Direction.teste = Weekly$Direction[!treino]
#Abaixo, temos a matriz de classificação
table(glm.pred2, Direction.teste)
```

```
##          Direction.teste
## glm.pred2 Down Up
##      Down    9  5
##      Up     34 56
```

```
#Total de Up
total_up2 = 5+56
total_up2
```

```
## [1] 61
```

```
#Total Down
total_down2 = 9+34
total_down2
```

```
## [1] 43
```

```r
#Sensibilidade (considerando UP como a classe positiva)

sensibilidade2 = 56/total_up2
#Abaixo, temos a taxa do modelo classificar como Up quando ele realmente é Up.
sensibilidade2
```

```
## [1] 0.9180328
```

```r
#Especificidade

especificidade2 = 9/total_down2
#Abaixo, temos a taxa do modelo classificar como Down quando ele realmente é Down.
especificidade2
```

```
## [1] 0.2093023
```

```r
#Total de predições corretas

total_correto2 = (9+56)/(total_down2 + total_up2)
#Abaixo, temos a fração total de predições corretas
total_correto2
```

```
## [1] 0.625
```