Regressão Logística

EXERCÍCIO 1

a) Modelo de regressão linear: sobreviveu em função de classe do passageiro, sexo, idade, nº de irmãos/esposos abordo do Titanic, nº de parentes/filhos abordo do Titanic, passagem e local de embarque.

```
##
## Call:
## lm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
##     Fare + Embarked, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06092 -0.21656 -0.08607  0.22393  1.00290
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.2638441  0.2753352   4.590 5.07e-06 ***
## Pclass2     -0.1437791  0.0453742  -3.169  0.00158 **
## Pclass3     -0.3346416  0.0425032  -7.873 1.01e-14 ***
## Sexmale     -0.5021838  0.0283859 -17.691  < 2e-16 ***
## Age         -0.0058328  0.0010819  -5.391 8.99e-08 ***
## SibSp       -0.0409771  0.0130682  -3.136  0.00177 **
## Parch       -0.0163464  0.0182360  -0.896  0.37029
## Fare         0.0003474  0.0003401   1.021  0.30732
## EmbarkedC   -0.1010612  0.2713371  -0.372  0.70964
## EmbarkedQ   -0.1028007  0.2741059  -0.375  0.70772
## EmbarkedS   -0.1702427  0.2710056  -0.628  0.53004
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3794 on 880 degrees of freedom
## Multiple R-squared:  0.3988, Adjusted R-squared:  0.3919
## F-statistic: 58.36 on 10 and 880 DF,  p-value: < 2.2e-16
```

O p-valor da estatística F aponta que a regressão apresentou significância estatística global. Com um $R^2 = 0,3984$, apenas $39,84\%$ das variações na variável dependente são explicadas pelo conjunto de variáveis explicativas.

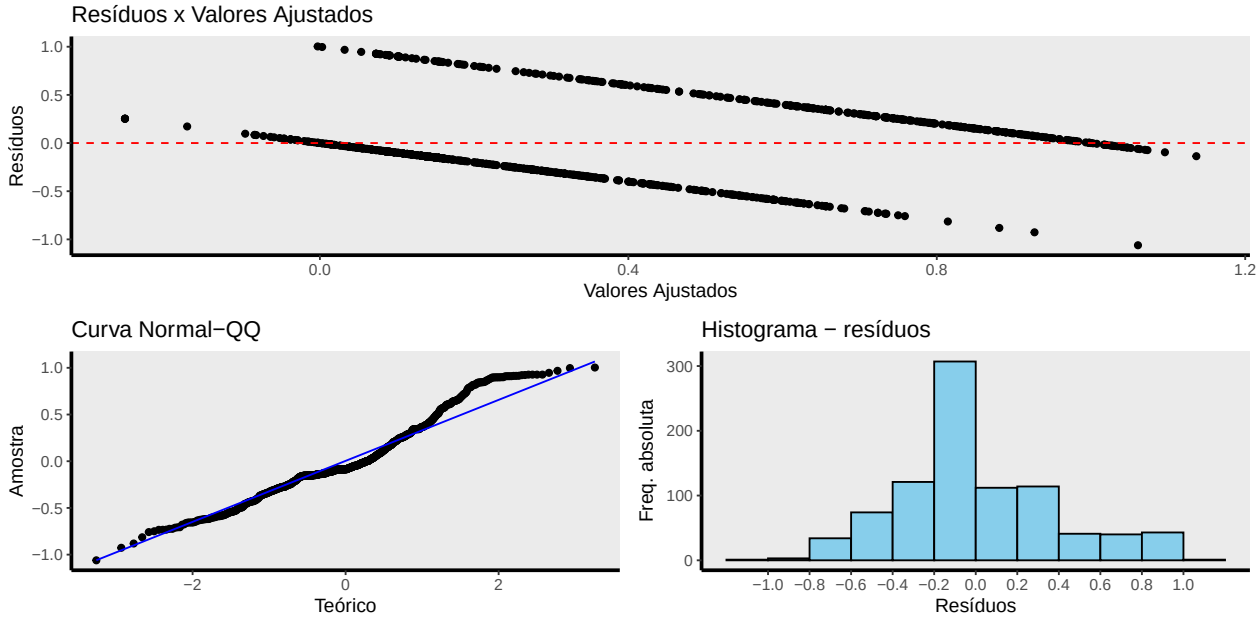
Em relação às significâncias individuais, temos que apenas a classe do passengeiro, seu sexo, idade e o número de irmãos/esposos no Titanic. Os resultados indicam que:

1. Passageiros da segunda classe apresentaram um valor $0,1438$ unidades menor para a variável dependente quando comparados aos passageiros da primeira classe;
2. Passageiros da terceira classe apresentaram um valor $0,3346$ unidades menor para a variável dependente quando comparados aos passageiros da primeira classe;
3. Passageiros do sexo masculino apresentaram um valor $0,5021$ unidades menor para a variável dependente quando comparados às passageiras;

4. O aumento de um ano na idade do passageiro esteve associado à uma redução de $0,0058$ unidades na variável dependente;

5. O aumento de uma unidade no número de irmãos/esposos do passageiro esteve associado à uma redução de $0,0410$ unidades na variável dependente.

Analisando agora os gráficos diagnósticos:



Claramente os resíduos são heterocedásticos, o que já era esperado, uma vez que a variável dependente é binária. Apesar disso, eles aparentam ser aproximadamente normais. Entretanto, o maior problema do modelo estimado (e dos modelos de probabilidade linear em geral) é a previsão de valores ajustados maiores do que 1 e menores do que 0, o que é impossível, pois por definição as probabilidades devem ficar entre 0 e 1. Isso pode ser visto abaixo, onde apresentamos a sumarização estatística dos valores previstos:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.2528  0.1256  0.3032  0.3838  0.6344  1.1366
```

Para corrigir isso, procedemos em seguida à estimação de um modelo de regressão logística.

b) Modelo de regressão logístico: sobreviveu em função de classe do passageiro, sexo, idade, nº de irmãos/esposos abordo do Titanic, nº de parentes/filhos abordo do Titanic, passagem e local de embarque.

Fazendo a estimação do modelo, obtemos a seguinte sumarização dos resultados

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
##     Fare + Embarked, family = "binomial", data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6235  -0.6098  -0.4222   0.6100   2.4512
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  16.414388 610.558089   0.027  0.97855
```

```
## Pclass2      -0.924047   0.297882  -3.102  0.00192 **
## Pclass3      -2.149626   0.297749  -7.220 5.21e-13 ***
## Sexmale      -2.709611   0.201336 -13.458  < 2e-16 ***
## Age          -0.039320   0.007888  -4.984 6.21e-07 ***
## SibSp        -0.322143   0.109545  -2.941  0.00327 **
## Parch        -0.095061   0.119028  -0.799  0.42450
## Fare          0.002261   0.002462   0.918  0.35842
## EmbarkedC   -12.311604 610.557974  -0.020  0.98391
## EmbarkedQ   -12.341443 610.558025  -0.020  0.98387
## EmbarkedS   -12.757357 610.557962  -0.021  0.98333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  783.74  on 880  degrees of freedom
## AIC: 805.74
##
## Number of Fisher Scoring iterations: 13
```

Veja que as mesmas variáveis continuam sendo significativas, inclusive apresentando os mesmos sinais, entretanto seus coeficientes são diferentes. Por completude, apresentamos abaixo a tabela de razões de chance do modelo.

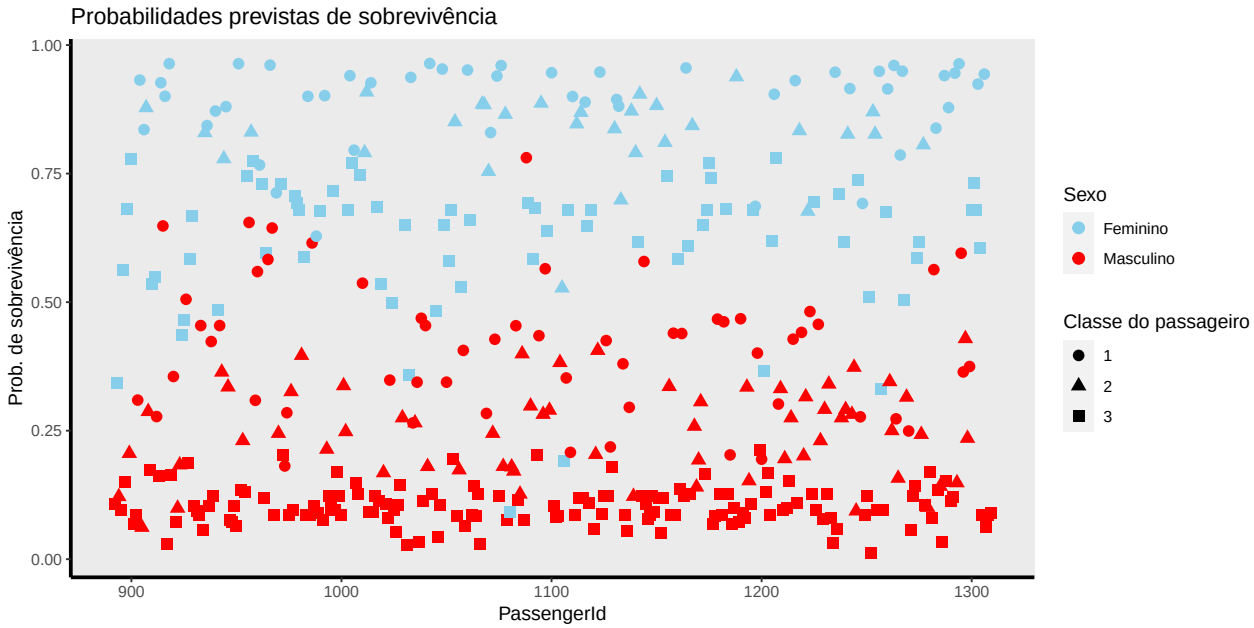
```
## Call:
## logitor(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
##     Fare + Embarked, data = dados)
##
## Odds Ratio:
##            OddsRatio  Std. Err.        z     P>|z|
## Pclass2   3.9691e-01 1.1823e-01  -3.1021  0.001922 **
## Pclass3   1.1653e-01 3.4696e-02  -7.2196 5.215e-13 ***
## Sexmale   6.6563e-02 1.3401e-02 -13.4582 < 2.2e-16 ***
## Age       9.6144e-01 7.5842e-03  -4.9845 6.213e-07 ***
## SibSp     7.2459e-01 7.9376e-02  -2.9407  0.003274 **
## Parch     9.0932e-01 1.0823e-01  -0.7986  0.424500
## Fare      1.0023e+00 2.4677e-03   0.9184  0.358424
## EmbarkedC 4.4992e-06 2.7470e-03  -0.0202  0.983912
## EmbarkedQ 4.3670e-06 2.6663e-03  -0.0202  0.983873
## EmbarkedS 2.8810e-06 1.7590e-03  -0.0209  0.983330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Em seguida, observando o sumário estatístico dos valores estimados,

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.006827 0.106420 0.272998 0.383838 0.676937 1.000000
```

podemos perceber que não mais ocorre a presença de probabilidades abaixo de 0 ou acima de 1, sendo está uma das razões que motiva o uso do modelo logístico quando a variável dependente é binária.

Para finalizar, temos abaixo o resultado da utilização do modelo estimado para prever a probabilidade de sobrevivência de uma amostra de 418 passageiros.



Fica claro que os passageiros que viajaram nas classes inferiores apresentaram menor probabilidade de sobrevivência, assim como os homens.