ACTL3142 Assignment PART 1

Kieren Chan z5478685

Executive Summary

This report provides an analysis on the health characteristics of patients in Brazil who have been deceased due to the Covid-19 pandemic. The data in this report is sourced from Hojo de Souza et al and aims to expand upon the authors work by exploring new datasets, such as ICU admission, to gain key new insights into the characteristics of covid death patients. Methods of analysis include graphical investigation, quantitative analysis and modelling our data into a generalised linear model. All higher order computations requiring R can be found in the appendix.

1 Exploratory Data Analysis

To begin, Exploratory Data Analysis (EDA) was conducted to help characterise patients who died from covid and find any commonalities or patterns.

1.1 Age and Gender

One point of interest that stems from our EDA was to find the ratio of deaths among the 2 genders and their ages.

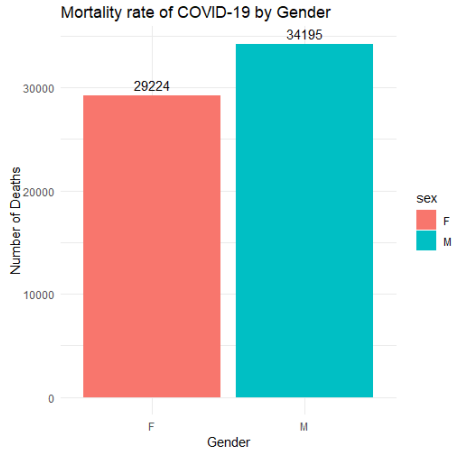

Figure 1: Bar plot of number of deaths per gender

It becomes apparent from figure 1 that the males have a greater mortality rate from covid by approximately 17%.

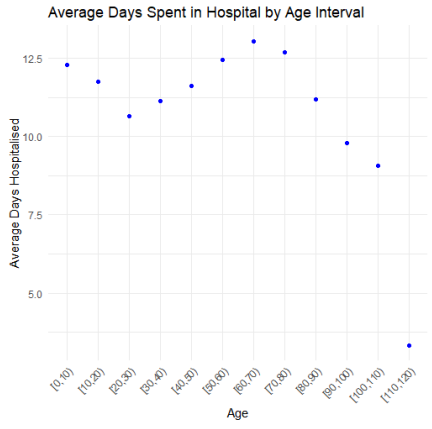

Figure 2: Scatterplot of Age vs Average days Hospitalised for Covid patients

Figure 2 portrays a graph with a shape like that of an inverted U, with a peak of 13.04 days and a trough of 9 days (ignoring the outlier of patients aged 110 and over). We can infer those patients in their 60's need relatively longer care suggesting their vulnerability to the virus. The spike in patients aged 10 and under can be accounted for because of their underdeveloped immune system at that age.

age_interval	deaths
<fct>	<int>
[0,10)	77
[10,20)	116
[20,30)	818
[30,40)	2661
[40,50)	6094
[50,60)	11291
[60,70)	15133
[70,80)	14554
[80,90)	9682
[90,100)	2870
[100,110)	121
[110,120)	2

Figure 3: Table of covid deaths per age interval of 10

We can further observe the vulnerability of patients aged 60 in figure 3, as we can see that the greatest number of deaths come from patients aged 60-70 in all, accounting for 23.86% of all deaths. Also considering that the median age in 2020 for Brazilians was 32.4 (O'Neill 2024), then the higher density of ages greater than 50 further supports that the older generations are more likely to die from Covid.

1.2 Vaccination Status and ICU

A reasonable area to explore in this EDA is the impact of vaccinations on the population.

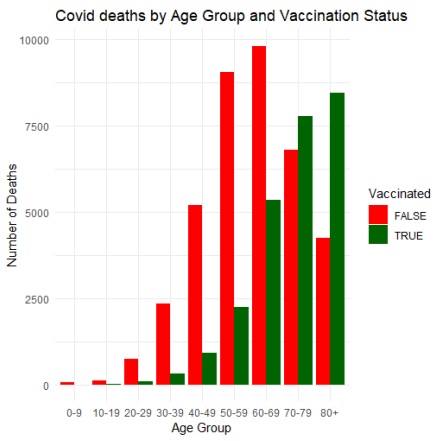

Figure 4: Bar graph of vaccination status by age vs Covid death

It is apparent from figure 4 that the vaccine has a drastic effect on covid death on those aged 69 and below. In comparison, those who received the vaccine aged 70 and above had more deaths, suggesting that the vaccine is less effective on older people aged above 70. It's also observed that as the age increases, the proportion of people who died with the vaccine in each age group increases exponentially.

	Survived	Died
Vaccinated	56,170	38,311
Unvaccinated	37,620	25,108

Table 1: Contingency table of survivors/deaths among vaccinated/unvaccinated patients

Additionally, there were in total 63419 deaths in all; approximately 25% of them were from patients aged 70 and above. We also see that 41% of unvaccinated patients went to ICU and 40.5% died inside ICU. These proportions are 39.9% and 40% respectively for vaccinated patients. Performing a Two Proportion Z Test with an alternative hypothesis being that the vaccinated proportions are less than the unvaccinated proportions, p values of 2.2×10^{-16} for ICU admission and 0.0194 and mortality are achieved. This means that vaccinated patients are less likely to be admitted to ICU and a lower mortality rate then the unvaccinated.

As a result, we can observe a gap of 17% in deaths between men and women, with men having more deaths at 34195. Moreover, individuals aged 60 or higher are more susceptible to becoming hospitalised most likely due to their deteriorating immune system quality, and the vaccine becomes less effective for many at that age. It is interesting to note that from our gatherings, we can infer men aged 60 and above are most vulnerable to this virus.

2 Model Fitting

Taking all our findings into account, we can create an improved model, compared to the full model, that more accurately models the characteristics of covid death. This section aims to explain and justify the reasoning behind the steps taken to create this improved model.

2.1 Creating the Improved Model

To improve the full model, a Logistical regression model was chosen as the covid death target variable is a binary (1,0), thus would be a logical selection. Next, a mixed stepwise selection function was used in R with the goal to minimise the Akaike Information Criterion (AIC). Using the full model as the input for the stepwise function, it was reconstructed such that parameters `sorethroat` and `vomit` were excluded while `DaysinHosp` was included, and `age` was converted into `dateBirth`.

2.2 Comparison of Full/Improved Model

To confirm that our new model is an improvement over the full model, we are required to assess multiple criteria based on goodness of fit statistics. Table 2 summarises these statistics, including AIC, Bayesian Information Criterion (BIC), deviance and log likelihood:

	Full	Improved
AIC	169,751.9	169,575.2
BIC	169,991	169,824.3
Deviance	169,704	169,691.41
Log-Likelihood	-84,852	-84,764

Table 2: Full/Improve model comparison

First, we achieve a new AIC of 169575.2 and a BIC of 169824.3 for our improved model, having a change in values of 176.7 and 166.7 respectively, which is a substantial change. A test of scaled deviance can be applied (with our significance level $\alpha = 5\%$):

$$D_1 - D_2 > 2(p - q).$$

$$12.59 > 4.$$

The ROC curve can also be constructed, with a classification threshold of 50%, and an Area under the Curve (AUC) value obtained as seen in figure 5.

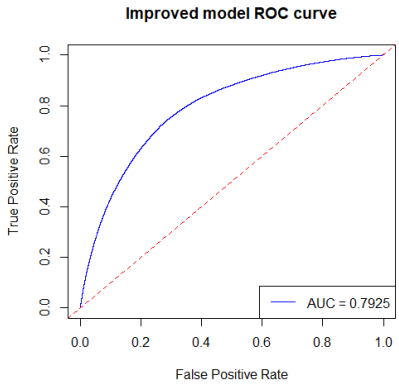

Figure 5: Improved model ROC curve

Despite yielding the same AUC result of 79%, the new model with fewer parameters and AIC suggest that much less of the noise was captured. This minimal change in AUC suggests that the parameters removed were just the noise parameters of the model and had little effect on the model, thus meaning that it was justified to remove such noise parameters. The replacement of the age predictor with `dateBirth` is also appropriate, not only does it decrease our AIC when replaced, but it provides us with a more precise value for the patient's age compared to having an integer result.

Therefore, our final model is an improvement over the full model by these statistical measures.

2.3 Findings of the Improved Model

After constructing our final model, we inevitably recognise its limitations. First, we realise that a GLM assumes a linear relationship between the outcome and predictors, which is not true as seen in figures 2 and 4. Also GLMs assume high multicollinearity among predictors however also not entirely true, especially when `sorethroat` and `vomit` were removed from because of their little influence on the model.

We also gained many insights, for one, our exclusion of the noise parameters justified because of their coefficients. For example, `vomitTRUE` suggested a patient who vomits has 0.011 higher log odds of dying from covid, especially insignificant with the high standard error with 0.02158. Thus, the noise parameters should have no issue with the patients for covid death. We can also refer to the coefficient of `dateBirth` being -2.9×10^{-4} constituting a log odd decrease in such magnitude of someone dying each day relative to our base case of 4th February 1907. This again aligns with our findings in section 1.1. The most impactful health condition to the patient is immunodeficiency, with increased log odds of dying at 0.61. This can be reasoned as patients with such condition have increased risk to diseases because of their complications.

Therefore, from these insights in our EDA and improved GLM, we gain an invaluable understanding of the characteristics of each patient. We can use this to help with aid and prevention to the covid pandemic and improve services and treatment available within the population of Brazil.

Appendix

A.1 Two Proportion Z Test for Vaccinated Vs. Unvaccinated ICU Admissions/Death

Our aim is to test the proportions of vaccinated patients admitted to ICU or die and see if it is less than the proportion of unvaccinated patients. This tests for the effectiveness of the vaccine in preventing intensive care or death. The Two Proportion Z test for our 2 cases:

Covid death test:

$$H_0: p_{Vac\ death} = p_{Unvac\ death} \quad vs \quad H_1: p_{Vac\ death} < p_{Unvac\ death}$$

ICU Test

$$H_0: p_{Vac\ ICU} = p_{Unvac\ ICU} \quad vs \quad H_0: p_{Vac\ ICU} < p_{Unvac\ ICU}$$

```
> vacICU_test <- prop.test(
   x = c(ICU_unvac, ICU_vac),
   n = c(unvaccinated, vaccinated),
   alternative = "less",
   correct = FALSE
)

> vacDeath_test <- prop.test(
   x = c(death_unvac, death_vac),
   n = c(unvaccinated, vaccinated),
   alternative = "less",
   correct = FALSE
)


> vacDeath_test

2-sample test for equality of proportions without continuity correction

data:  c(death_vac, death_unvac) out of c(vaccinated, unvaccinated)
X-squared = 4.2701, df = 1, p-value = 0.01939
alternative hypothesis: less
95 percent confidence interval:
 -1.000000000 -0.001066956
sample estimates:
   prop 1    prop 2
0.4002678 0.4054889


> vacICU_test

        2-sample test for equality of proportions without continuity correction

data:  c(ICU_vac, ICU_unvac) out of c(unvaccinated, unvaccinated)
X-squared = 4666.5, df = 1, p-value < 2.2e-16
alternative hypothesis: less
95 percent confidence interval:
 -1.000000 -0.145295
sample estimates:
   prop 1    prop 2
0.2651009 0.4139351
```

Because of the p-values of 0.01939 and 2.2×10^{-16} and chi-squared values being larger then p, for death and ICU admission tests respectively, we can safely reject our null hypotheses that vaccines give a lower chance of both ICU admission and dying, thus supporting our claim in section 1.2.

A.2 GLM Creation

A logistical model was used to model the full model (parameters such as `Patient_ID` was excluded as it clearly had no relation to their death status). The code below was used to create the full model:

```
# Changes TRUE covid death to 1 and false covid death to 0
data$covidDeath <- ifelse(data$covidDeath == 'TRUE', 1, 0)

# Create generalised linear model of full/"naive" model and our base improved
model
full_model <- glm(covidDeath ~ age + sex + vaccine + fever + cough + sorethroat
+ dyspnoea + oxygensat + diarrhea + vomit + hematologic + downsyn + asthma +
diabetes + neurological + pneumopathy + obesity + icu + respdistress + cardio +
hepatic + immuno + renal, data = data, family = binomial)
```

Then, to perform a stepwise selection, a new model `better_model` was created which included my EDA variables and `dateBirth`. The package `MASS` was used.

```
# Create new model that will be our improved model
scope <- list(
  lower = covidDeath ~ 1,
  upper = covidDeath ~ dateBirth + age + sex + vaccine + fever + cough +
sorethroat  +
    dyspnoea + oxygensat + diarrhea + vomit + asthma + diabetes +
    neurological + pneumopathy + obesity + icu + respdistress + cardio +
    hepatic + immuno + renal + hematologic + downsyn + DaysinHosp
)

best_model <- stepAIC(
  object = better_model,
  scope = scope,
  direction = "both",
)
```

A.3 GLM Summaries

Once the models were created, the `summary()` function was used to find the AIC, coefficients and standard errors of full/improved model.

```
> summary(full_model)

Call:
glm(formula = covidDeath ~ age + sex + vaccine + fever + cough +
    sorethroat + dyspnoea + oxygensat + diarrhea + vomit + hematologic +
    downsyn + asthma + diabetes + neurological + pneumopathy +
    obesity + icu + respdistress + cardio + hepatic + immuno +
    renal, family = binomial, data = data)

Coefficients:
                Estimate Std. Error  z value Pr(>|z|)
(Intercept)    -4.3169898  0.0362241 -119.174  < 2e-16 ***
age             0.0427408  0.0004539   94.158  < 2e-16 ***
sexM            0.1335948  0.0120094   11.124  < 2e-16 ***
vaccineTRUE    -0.5312894  0.0131877  -40.287  < 2e-16 ***
feverTRUE      -0.0311154  0.0124342   -2.502 0.012335 *
coughTRUE      -0.1552582  0.0136323  -11.389  < 2e-16 ***
sorethroatTRUE -0.0215803  0.0162338   -1.329 0.183733
dyspnoeaTRUE    0.1880748  0.0162112   11.602  < 2e-16 ***
oxygensatTRUE   0.2477212  0.0161799   15.310  < 2e-16 ***
diarrheaTRUE   -0.0447010  0.0175934   -2.541 0.011060 *
vomitTRUE       0.0109859  0.0215796    0.509 0.610691
```

```
hematologicTRUE   0.1083272  0.0628600    1.723 0.084833 .
downsynTRUE       0.4612620  0.0842329    5.476 4.35e-08 ***
asthmaTRUE       -0.1207874  0.0317375   -3.806 0.000141 ***
diabetesTRUE      0.1709797  0.0124880   13.692  < 2e-16 ***
neurologicalTRUE  0.4917597  0.0265381   18.530  < 2e-16 ***
pneumopathyTRUE   0.2676564  0.0283194    9.451  < 2e-16 ***
obesityTRUE       0.2091079  0.0158496   13.193  < 2e-16 ***
icuTRUE           1.8282585  0.0122642  149.072  < 2e-16 ***
respdistressTRUE  0.3157863  0.0138556   22.791  < 2e-16 ***
cardioTRUE        0.0162355  0.0123168    1.318 0.187448
hepaticTRUE       0.5302378  0.0551389    9.616  < 2e-16 ***
immunoTRUE        0.6153163  0.0340917   18.049  < 2e-16 ***
renalTRUE         0.5700748  0.0275329   20.705  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 212034  on 157208  degrees of freedom
Residual deviance: 169704  on 157185  degrees of freedom
AIC: 169752

Number of Fisher Scoring iterations: 4
```

Now checking the summary for the improved model.

```
Step:  AIC=169576
covidDeath ~ dateBirth + sex + vaccine + fever + cough + dyspnoea +
    oxygensat + diarrhea + hematologic + downsyn + asthma + diabetes +
    neurological + pneumopathy + obesity + icu + respdistress +
    cardio + hepatic + immuno + renal + DaysinHosp + age

               Df Deviance    AIC
<none>              169528 169576
+ sorethroat    1   169526 169576
- cardio        1   169530 169576
- hematologic   1   169531 169577
+ vomit         1   169528 169578
- dateBirth     1   169533 169579
- fever         1   169534 169580
- diarrhea      1   169535 169581
- asthma        1   169543 169589
- downsyn       1   169558 169604
- pneumopathy   1   169618 169664
- hepatic       1   169621 169667
- DaysinHosp    1   169641 169687
- sex           1   169654 169700
- dyspnoea      1   169662 169708
- cough         1   169665 169711
- obesity       1   169708 169754
- diabetes      1   169715 169761
- oxygensat     1   169768 169814
- immuno        1   169862 169908
- neurological  1   169875 169921
- renal         1   169970 170016
- respdistress  1   170043 170089
- vaccine       1   171203 171249
- age           1   179423 179469
- icu           1   193380 193426
```

A.4 ROC Curve and AUC

Both ROC curve and AUC computations were accomplished using the `pROC` library:

```
# Plot full model ROC curve
predicted_probs <- predict(full_model, type = "response")
roc_curve = roc(did_covidDeath, predicted_probs)
tpr <- roc_curve$sensitivities
fpr <- 1 - roc_curve$specificities

plot(fpr, tpr, type = "l",
     col = "blue",
     main = "Full model ROC curve",
     xlab = "False Positive Rate",
     ylab = "True Positive Rate",
     xlim = c(0,1))
abline(a = 0, b = 1, col = "red", lty = 2)
legend("bottomright", legend = paste("AUC =", round(auc(roc_curve), 4)), col =
"blue", lwd = 1)


# Plot improved model ROC curve
predicted_probs <- predict(best_model, type = "response")
roc_curve = roc(did_covidDeath, predicted_probs)
tpr2 <- roc_curve$sensitivities
fpr2 <- 1 - roc_curve$specificities

plot(fpr2, tpr2, type = "l",
     col = "blue",
     main = "Improved model ROC curve",
     xlab = "False Positive Rate",
     ylab = "True Positive Rate",
     xlim = c(0,1))
abline(a = 0, b = 1, col = "red", lty = 2)
legend("bottomright", legend = paste("AUC =", round(auc(roc_curve), 4)), col =
"blue", lwd = 1)
```

Generative AI Usage

ChatGPT was used to generate R code throughout the report for the purposes of error management, data analysis and model creation. See below for prompts:

- Error management
  - “[Pasted error code] Help me solve this error”
  - “How to omit NAN values”
- Data Analysis
  - “How to interpret stepAIC function in MASS library in R”
  - “How to create a contingency table in R”
  - “What does fdrop do”
  - “How to count how many females and males were tested”
- Model creation and analysis
  - “Logistic regression in R”
  - “How to add AUC value in ROC curve”

Bibliography

O'Neill, A. (2024) *Brazil - average age of the population 1950-2100*, *Statista*. Available at: https://www.statista.com/statistics/254361/average-age-of-the-population-in-brazil/ (Accessed: 27 June 2024).

Brazil (2024) *Worldometer*. Available at: https://www.worldometers.info/coronavirus/country/brazil/ (Accessed: 29 June 2024).

Garmendia, J. V., García, A. H., De Sanctis, C. V., Hajdúch, M. and De Sanctis, J. B. (2022) *Autoimmunity and Immunodeficiency in Severe SARS-CoV-2 Infection and Prolonged COVID-19*, *Current issues in molecular biology*. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9857622/#:~:text=Immunodeficiencies%20are%20strongly%20linked%20to,immune%20response%20is%20less%20probable. (Accessed: 30 June 2024).

(2023). Available at: https://www.thelancet.com/journals/lanam/article/PIIS2667-193X(23)00039-X/fulltext (Accessed: 30 June 2024).

de Souza, F. S., Hojo-Souza, N. S., Batista, B. D., da Silva, C. M. and Guidoni, D. L. (2021) *PLOS ONE*, 16(3). doi: 10.1371/journal.pone.0248580.