ACTL3142 Assignment Part I

Tadhg Xu-Glassop - z5480859

2024T2

Contents

1 Executive Summary **2**

2 Exploratory Data Analysis **2**
2.1 Vaccination Status . . . 2
2.2 Age and Mortality . . . 2
2.3 Days Spent in Hospital . . . 2

3 Generalised Linear Model Fitting **3**
3.1 Constructing the Model . . . 3
3.2 Comparing Final and Naive Models . . . 3
3.3 Insights from the Final Model . . . 4

4 Conclusion **4**

A Technical Appendix and AI Usage **5**
A.1 Two Proportion Z Test for Vaccinated vs. Unvaccinated ICU Admissions and Deaths . . . 5
A.2 t-Test for Mean Time Spent in Hospital of Survivors and Non-Survivors . . . 6
A.3 GLM Construction . . . 6
A.4 GLM Summaries . . . 7
A.5 GLM Information Criteria and Goodness-of-Fit Tests . . . 9
A.6 VIF of Predictors . . . 9
A.7 Confusion Matrix . . . 10
A.8 ROC and AUC Calculation . . . 10
A.9 Ordering of Final Model Coefficients . . . 10
A.10 Dependency of COVID Deaths . . . 11
A.11 Generative AI Usage . . . 11

1 Executive Summary

This report aims to build upon the discoveries of Hojo de Souza et al. in their study of COVID-19 and mortality risk by expanding upon their findings in a newer dataset with additional information, particularly vaccination status. Insights gained from this report will increase our understanding of key health characteristics that impact a patient's likelihood of dying to COVID, and ultimately aid in the assessment of the health risk and mortality of the Brazilian private health portfolio.

2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was initially performed on the `CovidHospDataBrasil.csv` dataset to gain a preliminary understanding of the profile of non-survivors of COVID-19 in the dataset. Resulting observations and findings will provide valuable insights into key factors that impact the likelihood of patients dying to COVID-19 and will provide foundational motivation for further exploration throughout this report.

2.1 Vaccination Status

A logical starting point of our EDA was to investigate the impact and effectiveness of the COVID-19 Vaccine in preventing hospitalisations, development of critical conditions (reflected by admission to the ICU), and ultimately death, which was unavailable in the Hojo de Souza et al. study[2].

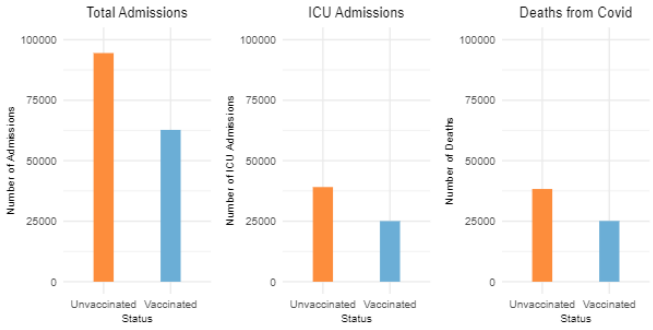

Figure 1: Vaccinated and Unvaccinated Patients, ICU Admissions and Deaths

Figure 1 shows the total number of admissions, ICU cases and deaths split into vaccinated and unvaccinated groups. We see a higher number of hospital admissions from cases of COVID-19 of unvaccinated patients in comparison to vaccinated, being 94,481 to 62,728 respectively. Combined with the context that in 2021, approximately 75% of the Brazilian population was vaccinated[5], this provides very strong evidence that those vaccinated are less likely to be administered to the hospital for cases of COVID-19.

In addition, we observe that of those administered to the hospital, 41.39% of unvaccinated patients were admitted to the ICU and 40.5% died, while these proportions are 39.9% and 40.0% for vaccinated patients. Performing a Two Proportion Z Test[1], we get p values of 3.66×10^{-8} and 0.0194 for ICU admission and mortality respectively. This provides very strong evidence that those vaccinated patients are less likely to develop critical cases and be admitted to the ICU, and also provides strong evidence that vaccinated patients have a lower mortality rate than unvaccinated patients.

2.2 Age and Mortality

A key finding in the study was that those most vulnerable to COVID-19 are of older generations. Thus, a very natural next step in our EDA was to investigate the age distribution of all patients and compare the different mortality rates between age groups.

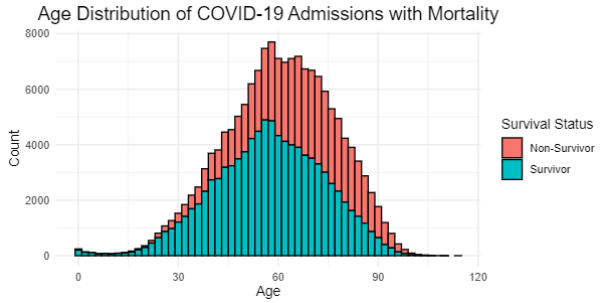

Figure 2: Age Distribution of Survivors and Non-Survivors

Figure 2 illustrates the age distribution of all patients admitted for COVID-19 and the proportion of survivors and non-survivors. Considering the fact that in 2020, the median age of the Brazilian population was 32.4 [7], the higher density of admissions in ages greater than 50 provides evidence that older generations are more likely to be hospitalised for cases of COVID-19 and are thus more vulnerable.

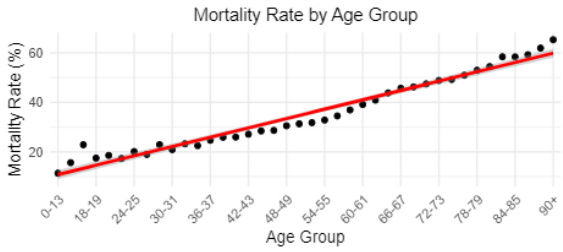

Figure 3: Mortality Rate of Hospitalised by Age Groups

Further, Figure 3 shows the sample mortality rate for each age group. We observe a somewhat linear trend in the mortality rate, with a low of 11.4% for the $0-13$ age group and a peak of 65.3% for those $90+$.

2.3 Days Spent in Hospital

A final statistic of interest was the number of days spent in hospital of survivors and non-survivors, which was not ad-

[1] See A.1 for technical details of hypothesis test

dressed in the study. Figure 4 shows the distribution of days spent in hospital of both survivors and non-survivors. It is clear from the plot that most survivors tend to have shorter stays with a median[2] stay of 7 days in comparison to non-survivors, who spend a median of 10 days in hospital before dying. A t-test also shows that the means of the two samples are statistically different[3], providing further evidence that hospitalisation length is a key predictor of mortality.

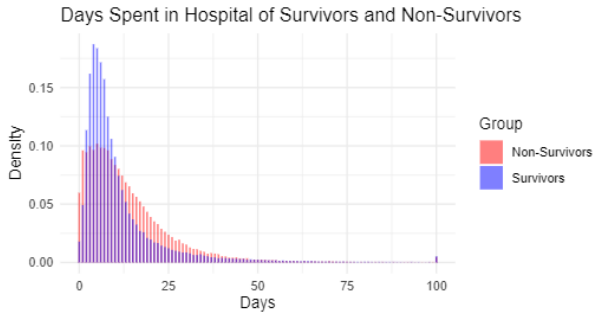

Figure 4: Total Days Spent in Hospital of Survivors and Non-Survivors

This is a result of more serious cases of COVID-19 having longer hospitalization periods due to significant underlying health characteristics, such as diabetes and age[3]. These patients required longer treatment plans and were observed to have a higher mortality rate, explaining our findings.

Thus, our EDA has provided us with strong evidence that vaccination status and age are key metrics of a patient's health characteristics that influence their likelihood of surviving COVID-19. In addition, we also found significant statistical differences in the hospitalisation length of survivors and non-survivors.

3 Generalised Linear Model Fitting

With an understanding of our dataset in mind, we move onto modelling the characteristics of COVID-19 hospitalisations that lead to death. This will lead to an improvement of our understanding of individual features and their relationships with each other, as well as corroborating and building upon findings from our EDA.

3.1 Constructing the Model

For this purpose, we use a *Logistic Regression Model* from the Generalised Linear Model (GLM) family. This specific model was chosen as the target variable `covidDeath` is Boolean (`TRUE` or `FALSE`), and thus Logistic Regression is appropriate as a classification model.

Initially, a *naive*[4] model was first constructed, which would act as a starting point for improvement until a sound model is reached. After performing hybrid step-wise[5] selection with consideration of AIC, BIC, and the addition of new variables, a final model with `sorethroat`, `vomit` and `cardio` excluded, the new variable `daysInHosp` included (time in days between hospital admittance and discharge or death) and the `age` variable replaced with `dateBirth` was reached[6].

3.2 Comparing Final and Naive Models

To confirm that our method in 3.1 did result in a sound model, we shall compare and assess the two models created - the *naive* model and the *final* model, to see if an improvement was made. This section aims to prove that the final model is indeed an improvement over the naive model.

First, we can compare information criteria and goodness-of-fit statistics of both models. Table 1 shows the results of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Deviance, and Log-likelihood measurement[7]. The improved model scores a preferred result in all 4 statistics, including the Log-Likelihood despite having 2 less predictors. Thus, by all these tests, the final model is preferred.

	Naive	Final
AIC	169,752	169,573
BIC	169,991	169,792
Deviance	169,704	169,529
Log-Likelihood	−84,852	−84,765

Table 1: Comparison of Naive and Final Models

Furthermore, we can consider a confusion matrix[8] with a classification threshold of 50%. The True Positive Rate (TPR) and False Positive Rate (FPR) of both models is summarised in Table 2. The final model is seen to have a higher TPR and a lower TPR and thus provides further evidence that our final model is an improvement of the naive model.

Model	TPR	FPR
Naive	0.6189387	0.1902122
Improved	0.6190176	0.1884529

Table 2: True Positive Rate and False Positive Rate of Naive and Final Models at 50% Threshold

We can observe the nature of both models at different thresholds with a Receiver Operating Characteristic (ROC) curve and comparing the Area Under Curve (AUC) statistics. Calculating the AUC for both models yields the same result of 0.79[9]. However, the final model having less param-

[2] Median is used instead of mean as the data is skewed by very long stays by some patients, as seen in Figure 4.
[3] See A.2 for details of hypothesis test.
[4] See A.3 for details of the naive model.
[5] See A.3 for technical details of model construction.
[6] For summaries of the naive and final model, see A.4.
[7] See A.5 for calculations of these.
[8] See A.7 for technical details about the confusion matrix.
[9] See A.8 for technical details of the ROC curve and AUC calculation, including a plot of the ROC curves.

eters but the same AUC as the naive model suggests the parameters removed in 3.1 were noise parameters and had little contribution to the model. Thus, this shows that the removal of the three parameters was justified and our model is an improvement as it is rid of noise parameters present in the naive model, that is, it is no longer over-fitting.

The final model also uses the `dateBirth` predictor as opposed to `age` predictor in the naive model. This change in predictor leads to a decrease of 66 in the AIC and BIC of the model[10] as a result of the increased accuracy from modelling patient's exact birth date, as opposed to an integer age. The use of this predictor also increases the applicability of the model, as each patient's profile can be more accurately depicted by the model.

Thus, it is clear by comparing information criterion, goodness-of-fit tests and the applicability of both models that the final model is an improvement upon the naive model. However, logistic regression requires the assumption that each observation is independent. This assumption is hard to justify, as a contagious virus such as COVID can have correlated deaths in local outbreaks and temporal dependencies from high infection and death rates in a previous period[11]. Therefore, although our final model is an improvement, we must address such limitations and be wary of potentially biased estimates and inferences.

3.3 Insights from the Final Model

After proving our final model is an improvement upon the naive model, we can gain insights from this model and the process we used to reach it.

First, the exclusion of `sorethroat`, `vomit` and `cardio` in the final model, in combination with the high p values of these variables in the naive model[12], show these symptoms have little significance on a patient's probability of surviving. For example, the coefficient of `vomitTRUE` in our naive model 0.01175 results in an interpretation that a patient who vomits has 0.01175 higher log-odds of dying, which is a relatively insignificant change and is further questioned by the high standard error of 0.02158. Thus, we find these 3 symptoms should cause little concern relative to other symptoms and health issues.

Now, we can build upon our findings from our EDA in section 2 by analysing predictors in our final model. Our finding that vaccination status impacts mortality risk of COVID in section 2.1 is supported by the model. The coefficient of `vaccineTRUE` being -0.53027 representing a decrease of 0.53027 in the log-odds of dying for a patient with the vaccine, with the low standard error of 0.01318 and it being the second largest coefficient in magnitude highlights the impact of the vaccination in preventing COVID fatalities. In addition, the coefficient of `dateBirth` -1.174×10^{-4} represents a decrease of the same magnitude in the log-odds of someone dying for each day they are younger than the base case of 4^{th} February, 1907, which is equivalent to -0.0429 for each year. This agrees with our findings in 2.2 and quantifies Figure 3, as older patients are more vulnerable and have higher mortality rate than younger ones. Similarly, the addition of `daysInHosp` to the final model quantifies our findings in section 2.3, that non-survivors tend to have longer stays in the hospital. The coefficient of -0.00477 represents the same decrease in the log-odds of surviving each day a patient spends in the hospital.

`icuTRUE` has the largest coefficient of 1.828, meaning patients admitted to the ICU have a significant 1.828 higher log-odds of dying. This quantifies a straightforward claim - that those admitted to the ICU tend to have much more serious cases of COVID than those who don't, and thus have a higher risk of dying. An interesting note is that the Variance Inflation Factor (VIF)[13] of `icuTRUE` is an insignificant 1.148, meaning that this predictor does not have significant correlation with other predictors and presents unique information.

By ordering the coefficients of our final model[14], we can gain an understanding of the severity of specific medical traits that higher risk patients possess. The most significant medical trait is immunodeficiency, with immunodeficient patients having 0.6261 higher log odds of dying. This is a result of such patients due to their increased risk of infection for other infections and complications, which in combination with COVID leads to a high mortality rate[1]. Out of these known upon admission, patients with neurological conditions have the highest risk, with 0.497 higher log odds of dying. While the connection is not fully understood, this corresponds with studies that have found patients with conditions such as dementia or Alzheimer's disease to be much more vulnerable to severe cases of COVID[6][9]. With these insights, we gain a better understanding of which patients are more vulnerable than others based off their underlying health conditions and the nature of their hospitalization.

4 Conclusion

Thus, our initial EDA and final GLM of the `CovidHospDataBrasil.csv` dataset has given us invaluable insights into the nature of COVID cases and the impact medical traits can have on the mortality of patients admitted for COVID-19. Throughout this report, we have built upon findings of de Souza et al.'s study and uncovered new insights with regards to vaccination status and hospitalization duration and their relation to COVID mortality. This information can be better used to assess the mortality and risk of the Brazilian health portfolio and improve services and responses to the COVID-19 Pandemic to provide better services and aid the Brazilian population.

[10] See A.3.

[11] See A.10 for a technical discussion and visualisation of the dependency of COVID deaths in the dataset.

[12] For all p values and coefficients of both models, see A.4.

[13] See A.6 for technical details of each VIF.

[14] See A.9 for more details.

A Technical Appendix and AI Usage

A.1 Two Proportion Z Test for Vaccinated vs. Unvaccinated ICU Admissions and Deaths

We aim to test whether the proportions of vaccinated patients admitted to the ICU or die is less than the proportion of unvaccinated patients to see if the vaccine is effective in preventing serious conditions and death against COVID. To do this, we will use the Two Proportion Z Tests for both cases, with the following hypothesis:

- For the "ICU Test,"

$$H_0 : p_{\text{Vac ICU}} = p_{\text{Unvac ICU}} \quad \text{v.s.} \quad H_1 : p_{\text{Vac ICU}} < p_{\text{Unvac ICU}}$$

- For the "Death Test,"

$$H_0 : p_{\text{Vac Death}} = p_{\text{Unvac Death}} \quad \text{v.s.} \quad H_1 : p_{\text{Vac Death}} < p_{\text{Unvac Death}}$$

We can perform the above tests with the following code, where the data frames `vac_death`, `num_vac`, and `vac_icu` contain the counts of admissions to the hospital and ICU and deaths of vaccinated and unvaccinated patients.

```
   # Perform the two-proportion z-test on the
> vac_death_test <- prop.test(
  x = c(vac_death[1,2], vac_death[2,2]),
  n = c(num_vac[1,2], num_vac[2,2]),
  alternative = "less",
  correct = FALSE
)

> vac_icu_test <- prop.test(
  x = c(vac_icu[1,2], vac_icu[2,2]),
  n = c(num_vac[1,2], num_vac[2,2]),
  alternative = "less",
  correct = FALSE
)

> vac_death_test
2-sample test for equality of proportions without continuity correction

data: c(vac_death[1, 2], vac_death[2, 2]) out of c(num_vac[1, 2], num_vac[2, 2])
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

> vac_icu_test

2-sample test for equality of proportions without continuity correction

data: c(vac_icu[1, 2], vac_icu[2, 2]) out of c(num_vac[1, 2], num_vac[2, 2])
X-squared = 33.449, df = 1, p-value = 3.658e-09
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.01048131
sample estimates:
  prop 1  prop 2
0.3992954 0.4139351
```

From the output, we observe our p values of 0.01939 and 3.658×10^{-9} for the death proportion and ICU admission proportion tests respectively, providing strong evidence that those with the vaccine have a lower chance of being admitted to the ICU and dying, supporting our findings in Section 2.1.

A.2 t-Test for Mean Time Spent in Hospital of Survivors and Non-Survivors

We aim to test whether the mean hospitalisation period of survivors and non-survivors is statistically different to gain further insights into factors that contribute to the likelihood of dying to COVID. To do this, we can perform a t test between the datasets of hospitalisation lengths of survivors and non-survivors with the following hypothesis, where μ is the population mean of hospitalization length,

$$H_0 : \mu_{\text{Survivors}} = \mu_{\text{Non-Survivors}} \quad \text{v.s.} \quad \mu_{\text{Survivors}} < \mu_{\text{Non-Survivors}}$$

The following code achieves this.

```
# perform t test for test of significant difference
> non_survivors <- data %>%
  filter(covidDeath)

> survivors <- data %>%
  filter(!covidDeath)

> t.test(
  x = survivors$numSymptoms, y = non_survivors$numSymptoms,
  alternative = "less",
  mu = 0
)
Welch Two Sample t-test

data: survivors$numSymptoms and non_survivors$numSymptoms
t = -39.2, df = 138092, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf -0.3479912
sample estimates:
mean of x mean of y
 5.036017 5.399249
```

We see a very small p value, thus providing very strong evidence that the mean time spent in hospital of survivors is less than that of non-survivors.

A.3 GLM Construction

As outlined in Section 3.1, a linear regression model was selected and a naive implementation was created as a starting point for improvement, including all non-trivial variables (trivial variables such as `Patient_ID` which clearly would not impact `covidDeath` were excluded) and without adding any new variables. The following code is what was used to produce the naive model.

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

Following this, the `stepAIC()` function from the `MASS` package was used to perform stepwise selection using AIC to improve upon the naive model. Stepwise selection was chosen over forward or backwards, as it combines both methods and tends to lead to a better result. The new variable `daysInHosp` was also added to the list of possible predictors as an attempt to improve upon the model by including findings discussed in Section 2.3.

```
>scope <- list(
  lower = covidDeath ~ 1,
  upper = covidDeath ~ age + sex + vaccine + fever + cough + sorethroat + hematologic +
```

```
  dyspnoea + oxygensat + diarrhea + vomit + asthma + diabetes + downsyn +
  neurological + pneumopathy + obesity + icu + respdistress + cardio +
  hepatic + immuno + renal + daysInHosp
)

>improved_model <- stepAIC(
   object = naive_model,
   scope = scope,
   direction = "both"
)
```

Finally, a model with all predictors in the current `final_model`, with `age` swapped for `dateBirth` was made. This was done manually, as adding `dateBirth` to `scope` above caused the function to include both predictors, leading to collinearity issues. These two models were then directly compared to arrive at the final model.

```
    # change out age for date of birth and compare the models
> model_with_dob <- glm(
   covidDeath ~ dateBirth + sex + vaccine + fever + cough + dyspnoea + oxygensat +
   diarrhea + asthma + diabetes + neurological + pneumopathy + obesity + icu +
   downsyn + hematologic + respdistress + hepatic + immuno + renal + daysInHosp,
   data = data,
   family = binomial()
)

> glance(improved_model)
# A tibble: 1 x 8
  null.deviance df.null logLik AIC BIC deviance df.residual nobs
         <dbl> <int> <dbl> <dbl> <dbl>  <dbl>    <int> <int>
1     212034. 157208 -84797. 169639. 169869. 169593. 157186 157209
> glance(model_with_dob)
# A tibble: 1 x 8
  null.deviance df.null logLik AIC BIC deviance df.residual nobs
         <dbl> <int> <dbl> <dbl> <dbl>  <dbl>    <int> <int>
1     212034. 157208 -84765. 169573. 169792. 169529. 157187 157209

# model with dob is better.
> improved_model <- model_with_dob
```

Other new variables, such as `timeInIcu` and `numSymptoms` were also tested, but would either result in collinearity issues signified by high VIF (see A.6) or a lower AIC when replacing correlated predictors. Hence, the only new variable was `daysInHosp`, as it did not appear to have significant correlation to other predictors in the dataset.

A.4 GLM Summaries

Coefficients and other measures of the naive model:

```
> summary(naive_model)

Call:
glm(formula = covidDeath ~ age + sex + vaccine + fever + cough +
    sorethroat + hematologic + dyspnoea + oxygensat + diarrhea +
    vomit + asthma + diabetes + downsyn + 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
hematologicTRUE 0.1083272 0.0628600 1.723 0.084833 .
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
asthmaTRUE   -0.1207874 0.0317375 -3.806 0.000141 ***
diabetesTRUE  0.1709797 0.0124880 13.692 < 2e-16 ***
downsynTRUE   0.4612620 0.0842329 5.476 4.35e-08 ***
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
```

Coefficients and other measures of the final model:

```
> summary(improved_model)

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

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.077e+00 2.272e-02 -91.440 < 2e-16 ***
dateBirth    -1.178e-04 1.220e-06 -96.585 < 2e-16 ***
sexM          1.352e-01 1.200e-02 11.262 < 2e-16 ***
vaccineTRUE  -5.283e-01 1.317e-02 -40.103 < 2e-16 ***
feverTRUE    -3.016e-02 1.235e-02 -2.441 0.014649 *
coughTRUE    -1.577e-01 1.350e-02 -11.681 < 2e-16 ***
dyspnoeaTRUE  1.879e-01 1.621e-02 11.589 < 2e-16 ***
oxygensatTRUE 2.513e-01 1.618e-02 15.529 < 2e-16 ***
diarrheaTRUE -4.467e-02 1.698e-02 -2.631 0.008508 **
asthmaTRUE   -1.209e-01 3.172e-02 -3.812 0.000138 ***
diabetesTRUE  1.711e-01 1.249e-02 13.699 < 2e-16 ***
neurologicalTRUE 4.930e-01 2.653e-02 18.582 < 2e-16 ***
pneumopathyTRUE 2.679e-01 2.835e-02 9.451 < 2e-16 ***
obesityTRUE   2.127e-01 1.584e-02 13.425 < 2e-16 ***
icuTRUE       1.866e+00 1.276e-02 146.212 < 2e-16 ***
downsynTRUE   4.671e-01 8.429e-02 5.542 2.99e-08 ***
hematologicTRUE 1.085e-01 6.295e-02 1.724 0.084778 .
```

```
respdistressTRUE 3.130e-01 1.383e-02 22.623 < 2e-16 ***
hepaticTRUE   5.331e-01 5.517e-02 9.662 < 2e-16 ***
immunoTRUE    6.224e-01 3.406e-02 18.272 < 2e-16 ***
renalTRUE     5.759e-01 2.755e-02 20.902 < 2e-16 ***
daysInHosp   -4.785e-03 4.516e-04 -10.597 < 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: 169529 on 157187 degrees of freedom
AIC: 169573

Number of Fisher Scoring iterations: 4
```

Note how the removed predictors `cardio`, `vomit` and `sorethroat` have the highest p values in the naive summary. In addition, the very low p value of `daysInHosp` further corroborates our finding in Section 2.3 that the time spent in hospital is related to the likelihood of dying and provides a unique insight that other predictors do not offer.

A.5 GLM Information Criteria and Goodness-of-Fit Tests

Information criteria and goodness-of-fit tests were used as an initial measurement of the improvement from the naive model to the final model due to the ease of getting such results and interpreting them. An interesting note briefly mentioned in Section 3.2 is that the log-likelihood test returned a higher (which is more favourable) result despite having 2 less predictors, when it is known that removing predictors will almost always result in a lower test result[4], thus providing evidence that the 3 removed predictors were indeed noise predictors and provided little benefit to the model.

Information criteria and goodness-of-fit test results seen in Table 1 were calculated with the `broom` package with the following code:

```
> glance(naive_model)
# A tibble: 1 x 8
  null.deviance df.null logLik AIC BIC deviance df.residual nobs
          <dbl> <int> <dbl> <dbl> <dbl>  <dbl>       <int> <int>
1      212034. 157208 -84852. 169752. 169991. 169704. 157185 157209

> glance(improved_model)
# A tibble: 1 x 8
  null.deviance df.null logLik AIC BIC deviance df.residual nobs
          <dbl> <int> <dbl> <dbl> <dbl>  <dbl>       <int> <int>
1      212034. 157208 -84797. 169639. 169869. 169593. 157186 157209
```

A.6 VIF of Predictors

VIF was calculated to check collinearity issues in the naive model and also to check if the addition of new variables would introduce any collinearity issues. The results below confirm that there were no such issues in either the naive nor the final model. As mentioned in A.3, some new variables were attempted to be introduced to the model to reduce AIC, but would cause collinearity issues signalled by a very significant VIF value. For example, when `numSymptoms` was introduced, it had a VIF of 70 and was thus removed to preserve the lack of collinearity issues.

VIF of predictors in both models were calculated using the `car` package with the following code:

```
> car::vif(improved_model)
dateBirth       sex    vaccine     fever     cough   dyspnoea oxygensat  diarrhea    asthma
1.369351  1.017974  1.181045  1.078329  1.067522  1.200851  1.184382  1.019082  1.012357
 obesity       icu    downsyn hematologic respdistress hepatic immuno    renal daysInHosp
1.109841  1.148283  1.008991  1.013374  1.216572  1.010473  1.018652  1.009774  1.089511
neurological pneumopathy diabetes
```

```
1.023034    1.019662    1.027320
```

A.7 Confusion Matrix

So far, only goodness of fit tests and indirect methods were used to compare the naive and final models. Thus, a confusion matrix was constructed to directly test the improvement in the final model's ability to predict. This was done by partitioning `CovidHospDataBrasil.csv` 80/20, where 80% of the data could be used to train new models using the naive and final predictors, and then tested on with the remaining 20% of the data, giving the confusion matrix. The test was then used to construct a confusion matrix with a threshold of 50%, as standard for binary target variables. Below is the actual confusion matrices for both models, which were used to calculate the TPR and FPR seen in Table 2. This act of direct testing shows the final model's improved ability in predicting, another facet of modelling that directly leverages the strengths of the model.

```
> naive_prediction_matrix

naive_predictions FALSE TRUE
          FALSE 15190 4833
          TRUE  3568 785

> improved_prediction_matrix

improved_predictions FALSE TRUE
            FALSE 15223 4832
            TRUE  3535 7851
```

A.8 ROC and AUC Calculation

Expanding upon A.7, we can visualise the difference in predicting performance between the two models with a ROC curve and consequently and AUC calculation. Below is the ROC curve, showing small differences in the models at different thresholds.

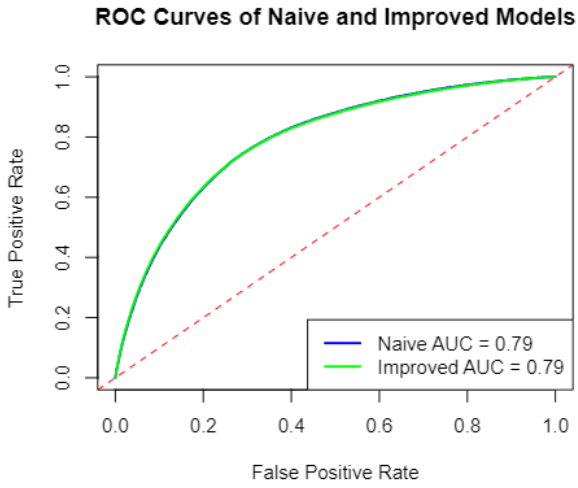

Figure 5: ROC Curve of Naive and Final Model with AUC

A.9 Ordering of Final Model Coefficients

By ordering the coefficients of the final model by *magnitude*, we get an understanding of total impact a predictor can have on the mortality rate of a patient. By taking the absolute value, we ignore whether a predictor increases or decreases the log-odds of a patient dying, but instead focus on the total change it has. By doing this, we see that admission to the ICU

by far has the biggest impact on the patients mortality, followed by prior hepatic conditions and vaccination status. This is contrasted by having a fever (ignoring `dateBirth` and `daysInHosp` as these are non-binary), which has a relatively small impact on log-odds of surviving if `TRUE`.

```
### Order coefficients by magnitude
> sort(abs(improved_model$coefficients))
  dateBirth   daysInHosp    feverTRUE  diarrheaTRUE hematologicTRUE  asthmaTRUE       sexM
 0.0001178211 0.0047853635 0.0301555077 0.0446665001 0.1084949747 0.1209141671 0.1351541381
  coughTRUE  diabetesTRUE  dyspnoeaTRUE  obesityTRUE oxygensatTRUE pneumopathyTRUE
  0.1577392252 0.1710516742  0.1879212881 0.2126733188 0.2512844508 0.2678839948
  respdistressTRUE downsynTRUE neurologicalTRUE vaccineTRUE hepaticTRUE renalTRUE immunoTRUE
  0.3129619053 0.4671102826 0.4930085294 0.5282673801 0.5331075444    0.5759459669 0.6223945246
   icuTRUE   (Intercept)
   1.8662390172 2.0772800316
```

A.10 Dependency of COVID Deaths

As mentioned in Section 3.2, our model requires the assumption that each observation of `covidDeath` is independent. This claim is hard to justify, and we can see the dependencies in the observations in Figure 6. We observe a low number of deaths at the start of 2021, which then increases exponentially to a peak in April 2021, caused by a major outbreak of COVID in Brazil[8], illustrating the temporal dependencies. Then, from July 2021, the number of COVID deaths decreases steadily due to the availability of the COVID vaccination in Brazil, showing a time dependent factor in our dataset.

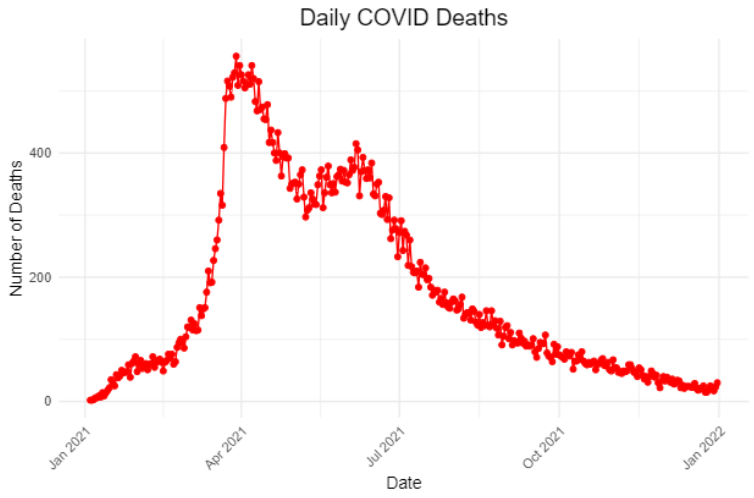

Figure 6: Daily COVID Deaths in the Dataset

A.11 Generative AI Usage

ChatGPT was used to generate `R` code for different purposes throughout the report, including plotting and cleaning data, modelling and analysing each model. Below are prompts used for each purpose.

- **Data Cleaning and Analysis**
  - "In my data, i have a boolean variable covidDeath and an integer column numSymptoms. i want to investigate the number of symptoms of TRUE and FALSE covidDeath. how can i do this"
  - "[Pasted some of my own code] here is some R code for a project im doing. at the start, ive added a new column that splits patients in my dataset into age groups. how can i add an age group that is just 96+?"
  - "if in the first half of my dataset there was no vaccination, and the second half in my dataset has the vaccination, then how does this impact dependability? (i observe a decrease in covid deaths)"

- **Data Visualisation and Plotting**
  - "how can i make a layered histogram (two variables plotted on the same dataset with lower alpha) in ggplot2?"
  - "without ggplot2, is there a way to change the thickness of the columns in base R? like in this function: barplot (copy and pasted my `barplot()` function)"
  - "how can i limit the number of x labels in ggplot2?"
  - "i have a dataframe non_survivors, with a dateEndObs representing the day they die. how can i plot deaths over time with this?"
- **Model Creation and Analysis**
  - "how do i use the bestglm function in R on my dataframe with the target variable covidDeath?"
  - "how do i use the stepAIC function from the MASS package to perform stepwise selection?"
  - "how can i get bic and adjr2 of my glm"
  - "i am trying to make a confusion matrix for two different models for comparison. my R code is: (copy and pasted my code) the table isnt giving me the matrix 2x2 matrix, but a massive list instead."
  - "say the coefficient of a high p value predictor in my logistic regression is 0.01175. this near 0 value means it is quite a negligible predictor right?"

References

[1] E Drzymalla et al. *COVID-19-related health outcomes in people with primary immunodeficiency*. Clin Immunol. Mar. 10, 2024. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9375253/.

[2] Hojo de Souza et al. *On the analysis of mortality risk factors for hospitalized COVID-19 patients: A data-driven study using the major Brazilian database*. PLoS ONE. URL: https://journals.plos.org/plosone/article/metrics?id=10.1371/journal.pone.0248580#citedHeader.

[3] Elzorkany K Shehab-Eldeen S Alarfaj HM Alarfaj SM Alabdulqader F Aldoughan A Agha M Ali SI Darwish E. Al Omair OA Essa A. *Factors Affecting Hospitalization Length and in-Hospital Death Due to COVID-19 Infection in Saudi Arabia: A Single-Center Retrospective Analysis*. PubMed Central. Jan. 8, 2023. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10404051/.

[4] J. Bruin. *How are the likelihood ratio, Wald, and Lagrange multiplier (score) tests different and/or similar?* Statistical Methods and Data Analytics. Jan. 2, 2023. URL: https://stats.oarc.ucla.edu/stata/ado/analysis/.

[5] J. Mendoza. *COVID-19 vaccine immunization development in Brazil 2021-2023*. statista. Apr. 10, 2023. URL: https://www.statista.com/statistics/1288019/population-vaccinated-against-covid-brazil/.

[6] NIH. *COVID-19 and the Nervous System*. National Institute of Neurological Disorders and Stroke. URL: https://www.ninds.nih.gov/current-research/coronavirus-and-ninds/covid-19-and-nervous-system.

[7] Aaron O'Neill. *World Population Prospects*. United Nations. URL: https://www.statista.com/statistics/254361/average-age-of-the-population-in-brazil/.

[8] da Silva Baum K. Sott MK Bender MS. *Covid-19 Outbreak in Brazil: Health, Social, Political, and Economic Implications*. Int J Health Serv. Apr. 10, 2022. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9445630/.

[9] Stephani Sutherland. *Long COVID Now Looks like a Neurological Disease, Helping Doctors to Focus Treatments*. Scientific American. Jan. 3, 2023. URL: https://www.scientificamerican.com/article/long-covid-now-looks-like-a-neurological-disease-helping-doctors-to-focus-treatments1/.