TECHNICAL STARTUP ADVICE

Gino Ferretti, Alessio Manfredi

14th July 2022

1 Introduction

A **technical consulting startup** which analyzes medical markets is carrying on *three* different projects for different clients.

HEART DISEASE PROJECT

2 Problem statement

The first project, which is commissioned from a US medical agency, brings the consulting startup to develop an analysis and an algorithm that intends to pinpoint the most relevant risk factors of coronary heart disease as well as predict the overall risk for it.

3 Dataset

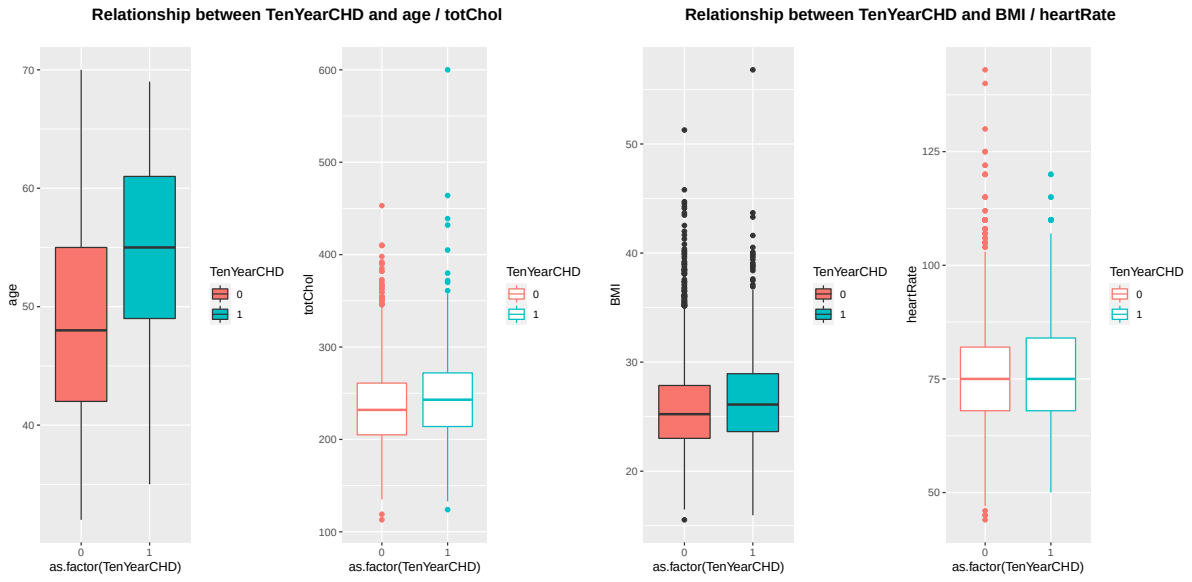
The data used are taken from the agency's database. The dataset is composed of 4238 observations and includes 16 variables (Table 1) based on patients' caratheristics. After cleaning it from *NA* values we were left with 3656 observations (no duplicates), which we then divided into two parts: the training set *(70% of the dataset)*, used to estimate the model, and the testing set *(the remaining 30%)*, used to assess the predictive accuracy of the model. Our target is TenYearCHD, a binary variable representing the 10-year risk of coronary heart disease; **3 variables describe some demographics** of the patients, **6 represent their current medical situation**, **4 represent the medical history** of each patient and **2 represent the patients' behaviour with respect to smoking**. We then carried an Exploratory Data Analysis of the training set and created boxplots to explore some variable relationships. In Figure 1 we plotted the variables of highest interest (*see rest of EDA in R script*), considering the purpose of the analysis.

To check for possible dependencies between the categorical covariates and the target variable we performed the Pearson's Chi-square test for **male** and **education**. After

Table 1: Variable description

Variable	Description	Type
male	gender (male or not)	character
age	age of the patient	integer
education	education level (1-4, asc)	integer
currentSmoker	current smoker or not	character
cigsPerDay	average smoked cigarettes per day	integer
BPMeds	blood pressure medication or not	integer
prevalentStroke	previous stroke or not	integer
prevalentHyp	being hypertensive or not	character
diabetes	being diabetic or not	character
totChol	total cholesterol level	integer
sysBP	systolic blood pressure	numeric
diaBp	diastolic blood pressure	numeric
BMI	Body Mass Index	numeric
heartRate	heart rate	integer
glucose	glucose level	integer
TenYearCHD	risk of CHD or not	integer

Figure 1: TenYearCHD and age/totChol/BMI/heartRate



running the test we rejected the null hypothesis of independence in both cases (*p-value = 0.0004082* for **male**, *p-value = 0.00002694* for **education**). Then we analyzed the correlation between all the possible variables and each other (*see R code for correlation plot*).

4 Method

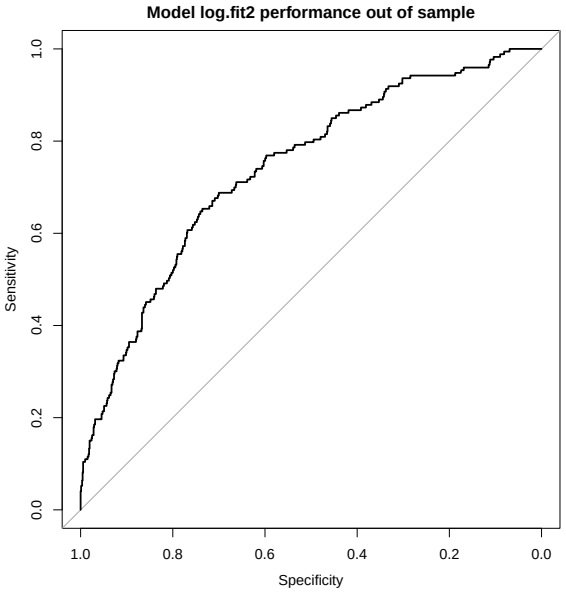
Considering the goal of the company we decided to construct the algorithm using a regression model in order to predict the overall risk of coronary heart disease for the patients. The models are based on the vector X containing in the first instance **all the variables**, and after observing the results the second model was deployed with the only the most relevant ones: X_i=**(male, age, cigsPerDay, sysBP, glucose)**. Since the variable of interest Y is binary, we decided to use a *logit model* (*glm* R function). We assume that $Y_i = TenYearCHD$ be a random variable with a *Bernoulli* distribution that is $Y_i = \mathbf{1}$ if the i-th patient is at risk ($p_i > 0.5$) and $Y_i = \mathbf{0}$ otherwise. The probability of CHD for unit i is p_i equal to the conditional expected value $E(Y_i|X_i)$.

5 Analysis

We used the testing set to evaluate the accuracy of the classification and the quality of the predicted probabilities. For the classification we obtained the **confusion matrix** and the corresponding *accuracy* of **84.87% for the first model** and **84.96% for the second one**, showing that the latter, with many fewer variables having the *highest relevance*, is more accurate than the former. To visualize performance at all possible thresholds of classification, we obtained the ***ROC curve*** shown in Figure 2 relative to the second model, with an AUC greater than the first one (*check R script for the latter*), confirming again our analysis.

Lastly, to assess the goodness of probabilities of our logistic regression once more, we performed the *Hosmer-Lemeshow* test manually. We divided the observations of the testing set into 10 similarly-sized groups (to overcome the size issue that the Pearson's chi-squared test could present) based on the deciles of the sorted expected probabilities (calculated using the model). We then obtained the number of both expected and observed successes and failures by summing up the predicted probabilities (expected successes). The final HL statistic is computed by summing all the $((obs-exp)^2/exp)$ for both successes and failures in each group, distributed as a *chi-square* distribution with $dF = groups-2 = 8$. The *p-value* for the first model is **0.758**, so we cannot reject the null hypotesis of equality between the observed and expected proportions, and same goes for the second model where the *p-value* is **0.736**. The results obtained by this analysis confirm one last time the good calibration of the models, according to the agency's goal. From the

Figure 2: Roc curve (AUC = 0.7398)



second model estimates (Table 2, all significant at 95% CI) we can see that, ceteris paribus, being a *male* increases the odds of having CHD in the next ten years by almost **40%** (*gender is the most impactful factor*). With every *year* the probability increases by **7%**, every additional *daily cigarette* brings it up by **2%**, 10 more units of *systolic blood pressure* bump up the odds by almost **16%**, and finally an increase in *glucose* levels of 10 leads to a greater probability of heart disease by nearly **6%**.

Table 2: Coefficient estimates of short regression of *TenYearCHD*

	Intercept	male	age	cigsPerDay	sysBP	glucose
Estimate	-8.485394	0.395486	0.070485	0.020229	0.016171	0.006333
Std.Err.	0.493140	0.127316	0.007638	0.005031	0.002501	0.001999

HEALTHCARE EXPENDITURE PROJECT

6 Problem Statement

The second project, commissioned from a hospital, challenges the consulting startup to perform an analysis and create an algorithm that predicts a range of possible healthcare expenses incurred by the patients.

7 Dataset

The data used for the analysis are offered by the hospital. The dataset is composed of 2955 observations and includes 8 variables (Table 3) based on some general characteristics of the patients, such as ID, total expenditure, number of chronic conditions, presence of additional insurance, and some demographics. The dataset presents no *NA* values, nor does it contain full row or ID duplicates. We split the sample in two groups: the training set *(60% of the observations extracted with SRS)*, used to estimate the model, and the testing set *(the remaining 40% of the dataset)*, used to assess the predictive accuracy of the model. The dependent variable to predict is totexp, which represents the total expenditure the patient sustains. In the whole dataset, the age of the patients is not normally distributed and ranges from 65 to 90 with a mean of 74.25. The overwhelming majority is white (97.36%) and 58.41% are female patients. The expenditure seems to have a left-skewed unimodal distribution, whereas its logarithm appears normally distributed. The simple linear regression of totexp on age seems to perform poorly as its diagnostic plots suggest heteroskedasticity and non-normality of the residuals. Using ltotexp we still have non-normality of the error, but now the residuals look homoskedastic.

Table 3: Variable description

Variable	Description	Type
dupersid	patient's ID	integer
totexp	patient's total expenditure (in thousands)	integer
ltotexp	natural logarithm of totexp	numeric
suppins	supplemental insurance or not	integer
totchr	number of chronic conditions	integer
age	patient's age	integer
female	female or not	integer
white	white or not	integer

8 Method

To fulfill the client's needs we explored the dataset and decided first to run a linear regression of totexp on all other variables ($\hat{totexp}_i = \hat{\beta}_0 + \hat{\beta}_1 suppins_i + \hat{\beta}_2 totchr_i + \hat{\beta}_3 age_i + \hat{\beta}_4 female_i + \hat{\beta}_5 white_i$) except for dupersid and ltotexp since we do not need the patient's ID as a regressor nor the logarithm of totexp if we are using the latter (or viceversa), then a quantile regression $\hat{q}^{QR}_{\tau,i} = x_i'\hat{\beta}_\tau$ for τ levels **0.025** and **0.975**. Subsequently we tried regressing ltotexp on all other variables (but the two explained earlier), followed by an *ANOVA test* on the coefficients of the first quantile regressions (totexp, for the two τ levels). By always referring to the totexp models, we created the indicator functions to

check whether an observation in the testing set is out of the 95% confidence interval (the quantile regression predictions outline those boundaries), and then proceeded to manually compute the likelihood-ratio test to check goodness of our predictions versus the actual outcomes. Finally, we explored the quantile regression plots of totexp (*see R code*) on a sequence of τ levels ranging from 0 to 1 by steps of 0.05 (both in sample and out of it), and decided that totchr was best represented, hence the final quantile regression of totexp on totchr only (same τ levels once again). We then repeated the likelihood ratio procedure to test the coverage of the new model out of sample.

9 Analysis

For the coefficient estimates of the regression models in this analysis section, please refer to Table 4: significance is assumed at a minimum standard 95% CI. The ***long regression of totexp*** returns significant estimates for totchr, female, and suppins, meaning that on average, ceteris paribus, ***healthcare expenditure increases with the number of chronic conditions, additional insurance, and if the patient is a male***. Diagnostic plots indicate that the residuals are heteroskedastic and not normal, so this is not a good model. For the ***quantile regression of totexp*** at $\tau_{0.025}$, **totchr** and **suppins** are significant, while at $\tau_{0.975}$ only **totchr** is. The long regression of ltotexp has significant intercept, totchr, suppins, and age estimators, with semi-normal residuals that are slightly heteroskedastic. The ***ANOVA test presents extremely-strong evidence (99.9%) to reject the null hypothesis of the coefficients in the two long quantile regressions of totexp at different τ being equal***. In order to create the indicator functions to check for failure of our models (actual observation falls out of boundary) we need to obtain the predictions of the quantiles for both models out of sample, which is simple to do for the quantile regression model since the predictions are the boundaries themselves, but for the linear regression model we have to get the predicted value (*expectation*) and then add/subtract the SE of the model multiplied by how many SE we need to reach the *0.025* and *0.975* quantiles of the residuals (which are not normal). Once we have done this we get the 95% CI for predictions of totexp with both models (linear and quantile), and we can create the indicator functions by assigning 1 if the observations falls out of bounds or 0 otherwise. We can therefore compute the log-likelihood function of this Bernoulli variable and use the formula $-2(log(model_{null}) - log(model_{alternative}))$, where the null model is for the ideal proportions (*2.5% of observations below and 2.5% above thresholds*) and the alternative one refers to the proportions we get with our predictions of the bounds (quantiles). The result of this test is distributed as a *chi-square* variable with $dF = dF_{alternative} - dF_{null} = 1$, and the *p-values we obtain are all bigger than 0.77* so **we do not have enough evidence to reject the null hypothesis of our models returning proportions equal to the actual ones (all of which means there is a**

good fit). When it comes to the quantile regression plots of *totexp*, we can clearly see how **totchr** seems to be the best predictor at many levels of quantiles (*check R script for plots and more*), which lead us to try a ***short quantile regression of totexp on totchr only*** (always in the training set) with the usual two τ levels of 0.025 and 0.975: **the estimates are all significant at both τ levels**. Finally, we repeated the same procedures to create the indicator functions, get the expected proportions of successes and failures, and calculate the likelihood-ratio test statistic for these Bernoulli variables, with final *p-values not smaller than 0.629*, indicating that the **shorter model with only one regressor performs as well as the more complicated one** and its predicted proportions are not statistically different from the real ones.

Table 4: Coefficient estimates for linear and quantile regressions of $\hat{Y}_i = \hat{totexp}_i$

	Intercept	suppins	totchr	age	female	white
long $LR(log(\hat{Y}_i))$	**5.515**	**0.332**	**0.456**	**0.016**	-0.053	0.343
long $LR(\hat{Y}_i)$	-1924.56	**1298.30**	**2739.66**	33.66	**-1288.86**	1795.35
long $qr_{\tau 0.025}$	-442.077	**170.923**	**235.769**	5.769	26.077	-123.923
long $qr_{\tau 0.975}$	58510.75	3448.125	**10737.719**	-609.406	-1640.281	2638.063
short $qr_{\tau 0.025}$	**27.667**	/	**200.333**	/	/	/
short $qr_{\tau 0.975}$	**18263.000**	/	**11396.250**	/	/	/

MODERNA'S VOLUME PROJECT

10 Problem Statement

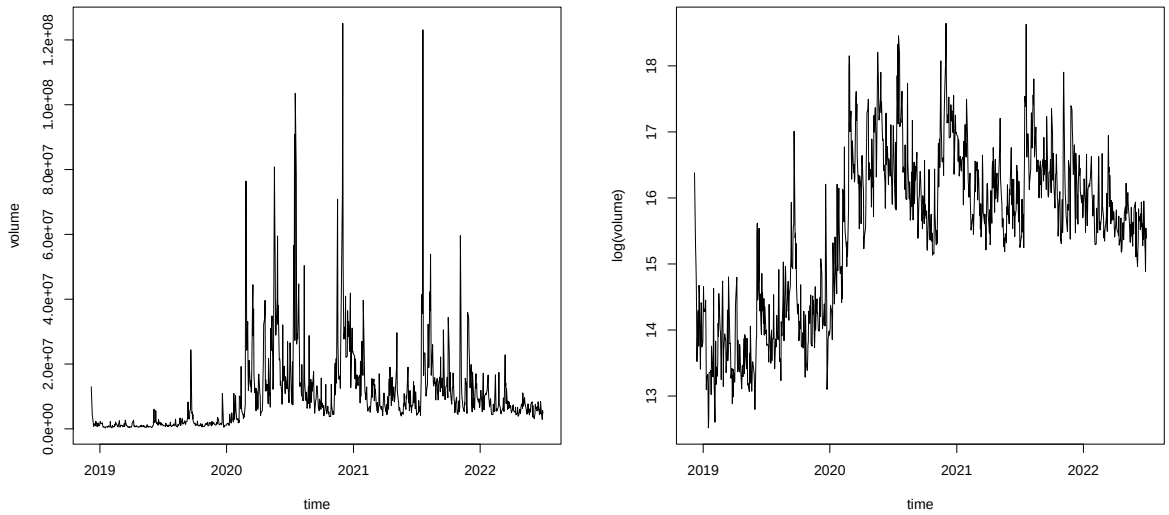
The third project, commissioned from a hedge fund specialized in medical markets, asks the consulting startup to analyze and design an algorithm that can forecast future values of Moderna's trading volume.

11 Dataset

The dataset to be explored is downloaded from a well-known financial reporting website, and it contains information about daily prices of shares (open/high/low/close/adjusted close) and volume from 2018 to 2022, for a total of 898 observations with neither NAs nor full-row/date duplicates. The sample was split into a training set (*70% of total*) and testing set (*the remaining 30%*) to measure model performance. Due to the given task, the only variable of interest is volume, on which we performed the *Box-Cox transformation* to try and stabilize variance. By generating the time series of the volume and of its natural

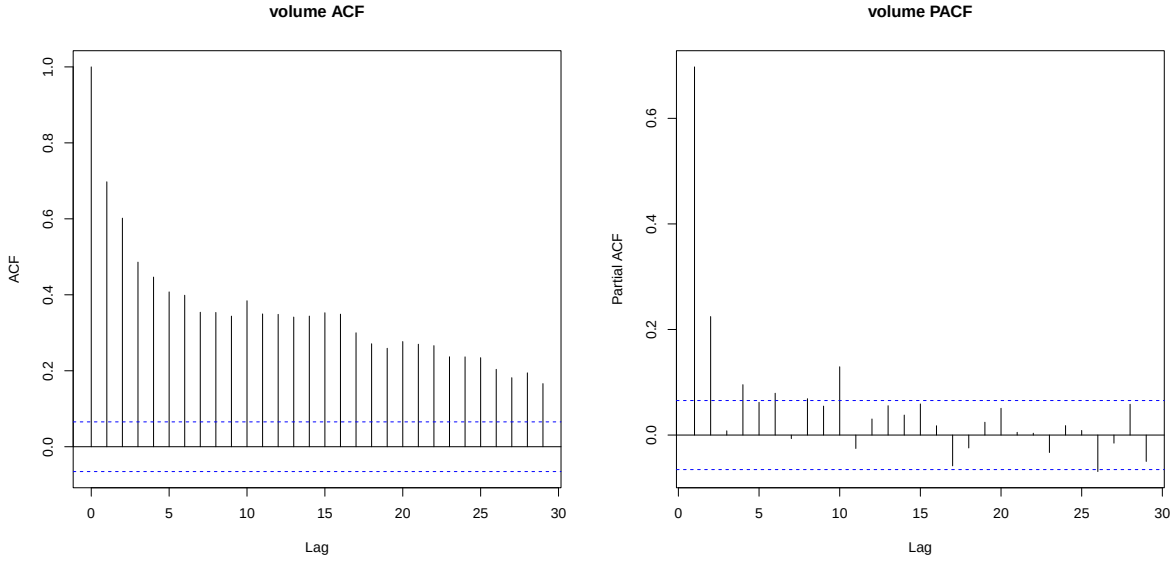
logarithm we obtained the plots in Figure 3, then we computed the ACF and the PACF (see Figure 4).

Figure 3: volume (left) and log(volume) (right) over time



The ACF of the volume decreases as the lag increases in a gradual way, with a considerable dip in the first few lags, typical of a stationary process. To test this, we proceed with an ***Augmented Dickey-Fuller*** test for the presence of a unit root in the *volume*, and a **p-value** $< \mathbf{0.01}$ supports our initial intuition of stationarity (no unit root), while another Augmented Dickey-Fuller test on *log(volume)* shows a **p-value** $= \mathbf{0.2783}$ which fails to reject non-stationarity; in order to get a stationary log(volume) series, it is necessary to differentiate it at least once, in which case the ADF test returns stationarity.

Figure 4: ACF and PACF of volume



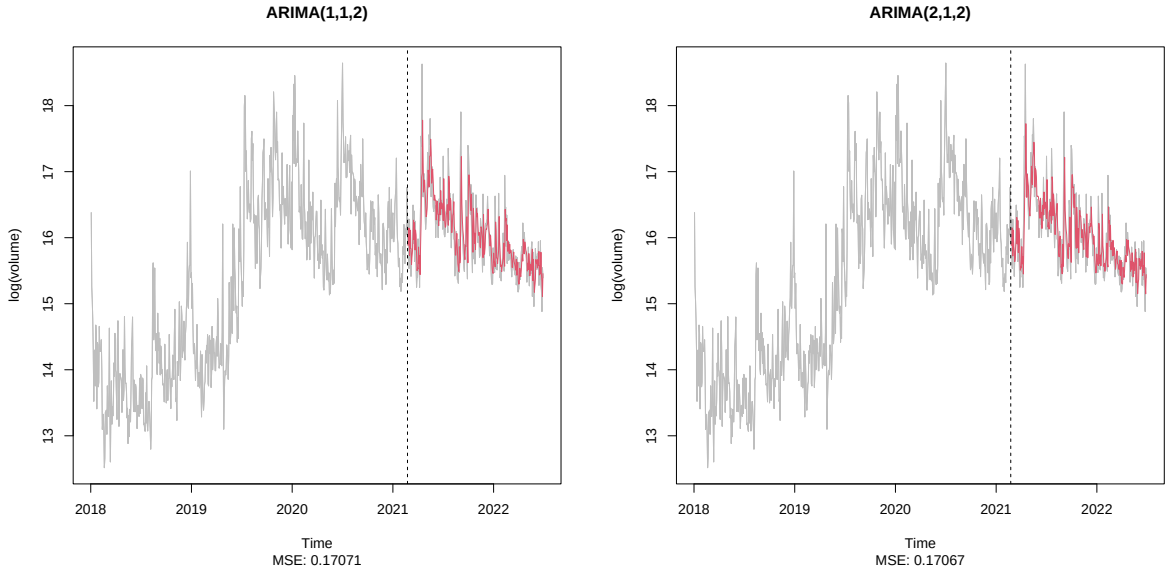
12 Method

To satisfy the client's request we investigated the dataset and tried to fit linear and quantile regression models on *log(volume)*: all performed very poorly (*see R script for full EDA and models as well as plots*) which could have been expected as the linear regression models (quantiles included) work better when the series is stationary (*log(volume)* appears not to be) or linearly-trend-stationary, which does not seem to be the case unless we suspect a *structural break* in the parameters of the linear regression (visible at around 2020 in the plots). Knowing that the linear models are not ideal and seeing how the *differentiated log(volume)* is stationary, the next logical step is to consider an *ARIMA* model that would automatically differentiate *log(volume)* on top of providing an *ARMA* model which deals much better with a stationary time series.

13 Analysis

Forecasts for *short term* (Figure 5) and *long term* (Figure 6) were made. We chose the best-performing models among the ones tried (*see R script for more insight into other models such as AR, MA, ARMA, and for missing ARIMA plots*): ***ARIMA(1,1,1), ARIMA(1,1,2), ARIMA(2,1,1), ARIMA(2,1,2)*** were picked out. The model classification table (*see Table 5*) was compiled using the *AIC* and the *MSE*.

Figure 5: ARIMA(1,1,2) and ARIMA(2,1,2) for short term



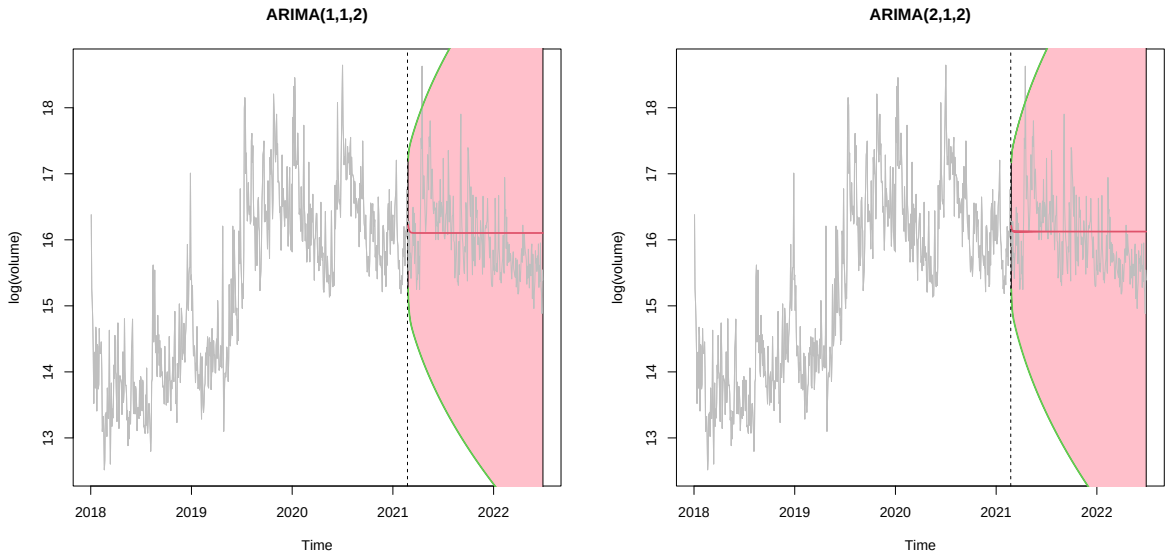
Observing the short-term forecast plots, the ARIMA models predictions all look similar to each other with a bit of an exception for *ARIMA (2,1,2)* which seems to cover more data. Concerning the long run, there is a more noticeable difference in the prediction intervals, despite it being rather small. Regardless of the forecast horizon, the four

proposed solutions seem to perform well without significant differences overall, meaning that all of them can be employed to develop accurate forecasts. However, ***we suggest adopting ARIMA(1,1,1) under information/time constraints (especially when working with big datasets), or the best-performing ARIMA(2,1,2) otherwise***, given how similar their scores are in Table 5.

Table 5: MSE and AIC model classification (ordered by best AIC)

	ARIMA(2,1,2)	ARIMA(1,1,1)	ARIMA(1,1,2)	ARIMA(2,1,1)
MSE	0.17067	0.17175	0.17071	0.17091
AIC	894.53	894.98	895.61	895.97

Figure 6: ARIMA(1,1,2) and ARIMA(2,1,2) for long term



Note that *ARIMA(2,1,2)* has slightly-larger intervals than *ARIMA(1,1,2)*, accounting for more variance. With these models, the client's needs are going to be met, with the assumption of the models being deployed on data of the same *type/nature* in order to produce future forecasts for the trading volume of interest (*Moderna Inc.*). As a final remark, the shares that were analyzed are from a company that has been listed on public markets for *only three years*, and it would probably be wiser to **target a company with more available data over a higher number of years**: this would help develop *more accurate models* and *better forecasts*.

References

1st project data set:***Heart disease data***
2nd project data set:***Healthcare expenditure data***
3rd project data set:***Moderna's shares data***