Coronavirus COVID-19 Probability Analysis

Fred Viole

3/17/2020

Probabilities of Death from COVID-19

Install latest NNS version (>= 0.5.0)

`NNS (0.5.0)` is available on CRAN via `install.packages("NNS")`.

```
# You need NNS 5.0
library(NNS)
```

Download Data & Read Data

This dataset is available on the data science website, Kaggle.

Last update: 03/13/2020, 8:00 PM (EST)

https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset#COVID19_line_list_data.csv

```
corona <- read.csv("COVID19_line_list_data.csv",header = TRUE)
```

Select Variables of interest

Will use `country`, `gender` and `age`.

We also create a variable `incubation_proxy` to serve as a proxy variable for virus incubation. We are using the length of exposure in number of days, but unfortunately, this reduces the number of complete observations to 88.

```
library(anytime)
incubation_proxy <- anydate(corona$exposure_end) - anydate(corona$exposure_start)

CV_model <- corona[, c("death", "country", "gender", "age")]


# Eliminate N/A values
CV_model <- CV_model[complete.cases(CV_model),]

# Some dates in death column...
CV_model$death <- ifelse(CV_model$death==0,0,1)

head(CV_model)
```

```
##   death country gender age
## 1     0   China   male  66
## 2     0   China female  56
## 3     0   China   male  46
## 4     0   China female  60
## 5     0   China   male  58
## 6     0   China female  44
```

Countries Available with this dataset

```
levels(CV_model$country)
```

```
##  [1] "Afghanistan" "Algeria"     "Australia"   "Austria"     "Bahrain"    
##  [6] "Belgium"     "Cambodia"    "Canada"      "China"       "Croatia"    
## [11] "Egypt"       "Finland"     "France"      "Germany"     "Hong Kong"  
## [16] "India"       "Iran"        "Israel"      "Italy"       "Japan"      
## [21] "Kuwait"      "Lebanon"     "Malaysia"    "Nepal"       "Phillipines"
## [26] "Russia"      "Singapore"   "South Korea" "Spain"       "Sri Lanka"  
## [31] "Sweden"      "Switzerland" "Taiwan"      "Thailand"    "UAE"        
## [36] "UK"          "USA"         "Vietnam"
```

Probability of Death Overall

```
mean(CV_model$death)
```

```
## [1] 0.07030303
```

Probability of Death in China

```
mean(CV_model[CV_model$country=="China", "death"])
```

```
## [1] 0.2052632
```

Probability of Death for >65 in China

```
mean(CV_model[CV_model$country=="China" & CV_model$age>=65, "death"])
```

```
## [1] 0.6666667
```

Variables of Interest

Select a specific `country`, `gender` and `age_of_interest` to isolate that specific probability.

```
country_of_interest = "China"
gender_of_interest = "male"
ages = 1:99
age_of_interest = 50


# For Multiple Patients
patients = cbind.data.frame(country = as.factor(rep(country_of_interest, 99)),
                            gender = as.factor(rep(gender_of_interest, 99)),
                            age = ages)
```

Baseline logit Model

```
logit = glm(death ~ country + gender + age, family = binomial, data = CV_model)
summary(logit)
```

```
##
## Call:
## glm(formula = death ~ country + gender + age, family = binomial,
##     data = CV_model)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7392  -0.2301  -0.0870   0.0000   3.5713
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.674e+01  4.065e+03  -0.007  0.99475
## countryCambodia    -1.390e+00  1.819e+04   0.000  0.99994
## countryCanada       5.140e-01  6.204e+03   0.000  0.99993
## countryChina        1.851e+01  4.065e+03   0.005  0.99637
## countryFinland      2.740e+00  1.819e+04   0.000  0.99988
## countryFrance       1.773e+01  4.065e+03   0.004  0.99652
## countryGermany      2.341e-01  5.954e+03   0.000  0.99997
## countryHong Kong    1.505e+01  4.065e+03   0.004  0.99705
## countryItaly        1.290e+00  1.819e+04   0.000  0.99994
## countryJapan        1.480e+01  4.065e+03   0.004  0.99710
## countryLebanon      1.347e+00  1.819e+04   0.000  0.99994
## countryMalaysia    -8.835e-02  5.144e+03   0.000  0.99999
## countryNepal        1.611e+00  1.819e+04   0.000  0.99993
## countryPhillipines  2.040e+01  4.065e+03   0.005  0.99600
## countrySingapore   -2.221e-02  4.396e+03   0.000  1.00000
## countrySouth Korea  1.776e+01  4.065e+03   0.004  0.99651
## countrySpain       -2.151e-01  5.666e+03   0.000  0.99997
## countrySri Lanka    1.883e+00  1.819e+04   0.000  0.99992
## countrySweden       3.490e+00  1.819e+04   0.000  0.99985
## countrySwitzerland -2.461e+00  1.819e+04   0.000  0.99989
## countryTaiwan       1.600e+01  4.065e+03   0.004  0.99686
## countryThailand    -7.459e-01  5.631e+03   0.000  0.99989
## countryUAE         -7.613e-01  7.094e+03   0.000  0.99991
## countryUK          -6.394e-01  1.819e+04   0.000  0.99997
## countryUSA         -5.777e-01  7.827e+03   0.000  0.99994
## countryVietnam      6.662e-02  6.713e+03   0.000  0.99999
## gendermale          1.129e+00  3.906e-01   2.891  0.00384 **
## age                 1.072e-01  1.426e-02   7.514 5.72e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 419.80  on 824  degrees of freedom
## Residual deviance: 230.31  on 797  degrees of freedom
## AIC: 286.31
##
## Number of Fisher Scoring iterations: 19
```

Overall `NNS` model

```
NNS_model = NNS.stack(IVs.train = CV_model[,2:4],
                          DV.train = CV_model$death,
                          order = "max", method = 1)$stack

# Ensure within [0,1]
NNS_model = pmax(0, pmin(NNS_model, 1))
```

Probability of Death for ages 1:99 given Country and gender

logit Model

```
logit_predictions = predict(logit, newdata = patients, type = 'response')
```

NNS Model

This is where the specific data of interest `patients` is used in the `IVs.test =` argument.

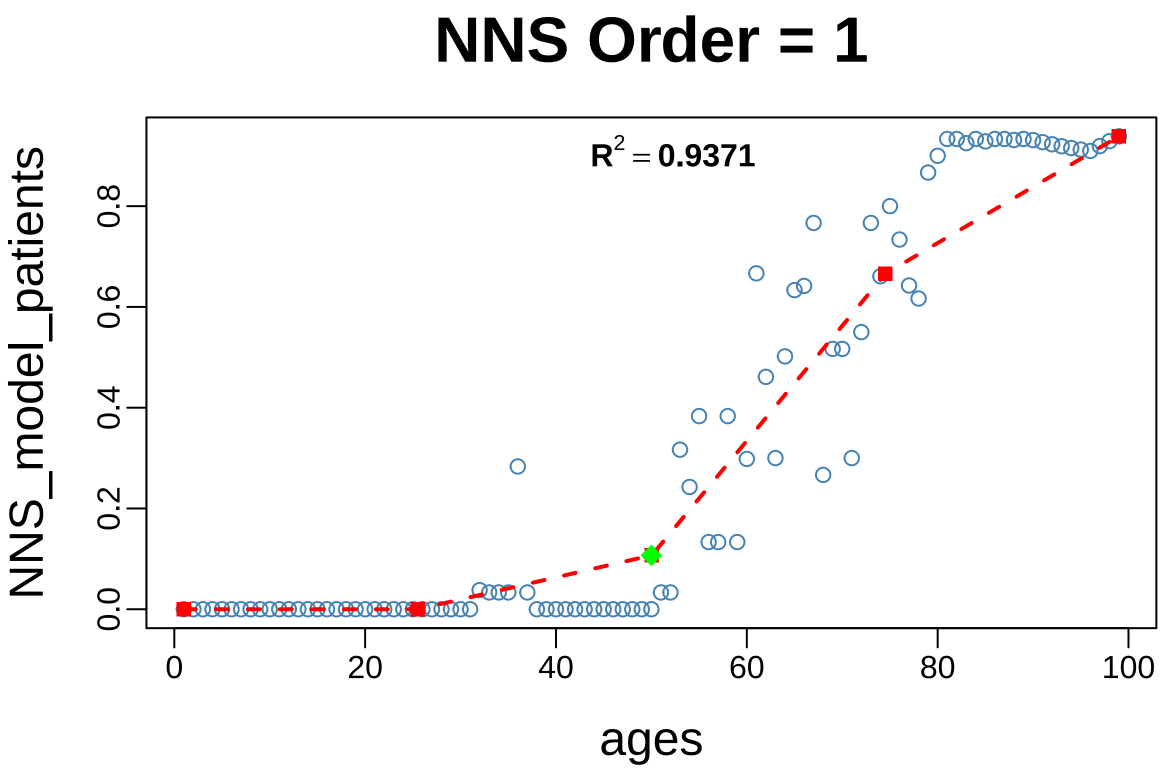
```
NNS_model_patients = NNS.stack(IVs.train = CV_model[,2:4],
                               DV.train = CV_model$death,
                               order = "max", method = 1,
                               IVs.test = patients)$stack

# Ensure within [0,1]
NNS_model_patients = pmax(0, pmin(NNS_model_patients, 1))
```

Regress Predictions on Age

We also isolate the `age_of_interest` from the point estimates for a specific probability of a given `age`, conditional on `country` and `gender` already provided.

```
NNS_predictions = NNS.reg(ages, NNS_model_patients, point.est = age_of_interest)
```



```
# Specific point estimate for given age
NNS_predictions$Point.est
```

```
## [1] 0.1069602
```

There is obviously a small sample issue at play here. For instance, there were 0 deaths for those males aged 50 in China, but the probability increases to 50% for those males aged 53 years. Is there reason to believe the probability of death would follow such a discrete step function for specific ages? No…

```
CV_model[CV_model$country=="China" & CV_model$age==50,]
```

```
##     death country gender age
## 17      0   China   male  50
## 45      0   China   male  50
## 66      0   China female  50
## 72      0   China female  50
## 174     0   China   male  50
## 185     0   China   male  50
```

Only 2 observations for those aged 53 in China…

```
CV_model[CV_model$country=="China" & CV_model$age==53,]
```

```
##    death country gender age
## 62     1   China   male  53
## 80     0   China   male  53
```

np model

np handles the discrete nature of the probabilities smoothly, yet all of the estimates are roughly in agreement with similar overall probabilities (actual death numbers in black, logit fitted values in green, np fitted values in blue and NNS fitted values in red) in the first plot and similar local slopes in the second plot.

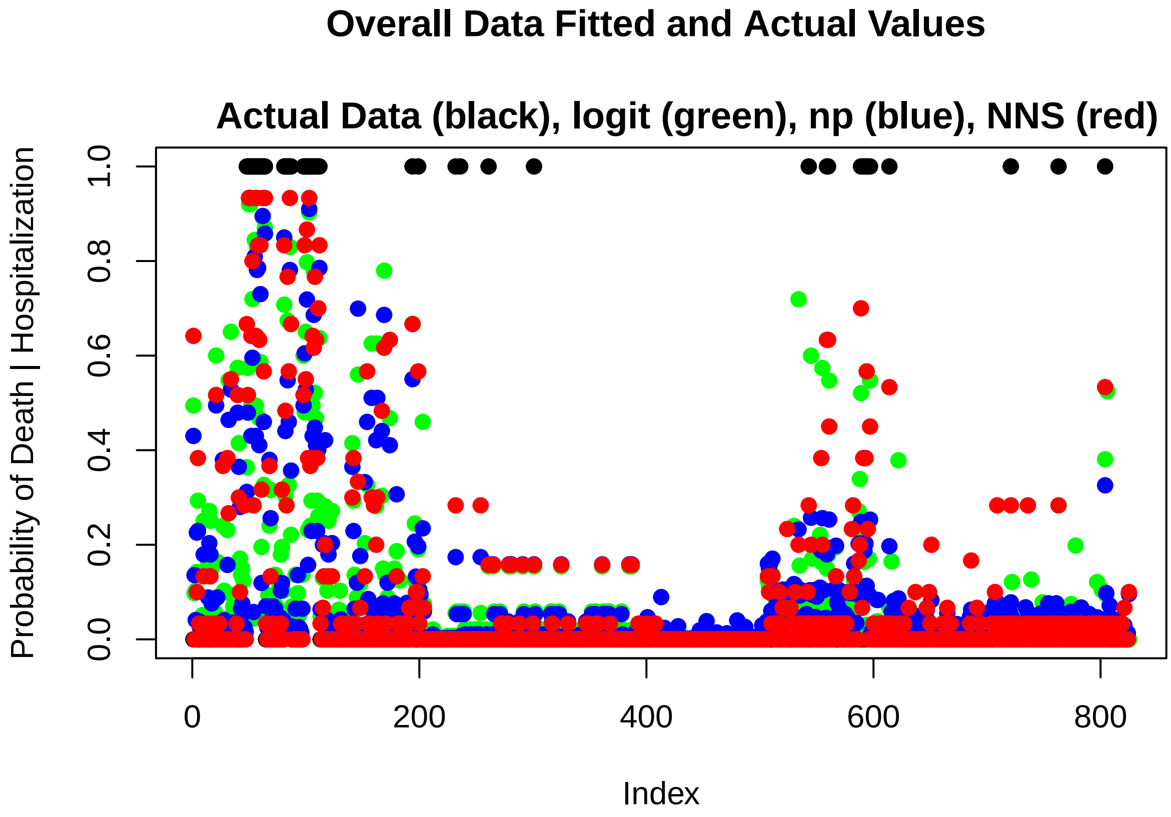
```
library(np)
```

```
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-10)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]
```

```
np_model = npreg(death ~ country + gender + age, data = CV_model)
np_predictions = predict(np_model, newdata = patients)
np_predictions[50]
```

```
## [1] 0.07585755
```

```
plot(CV_model$death, pch = 19, main = "Overall Data Fitted and Actual Values \n
     Actual Data (black), logit (green), np (blue), NNS (red)",
     ylab = "Probability of Death | Hospitalization")
points(logit$fitted.values, pch = 19, col = 'green')
points(np_model$mean, pch = 19, col = 'blue')
points(NNS_model, pch = 19, col='red')
```



Note the spikes at 100 and 600, which reflects China and Taiwan data:

```
CV_model[100, "country"]
```

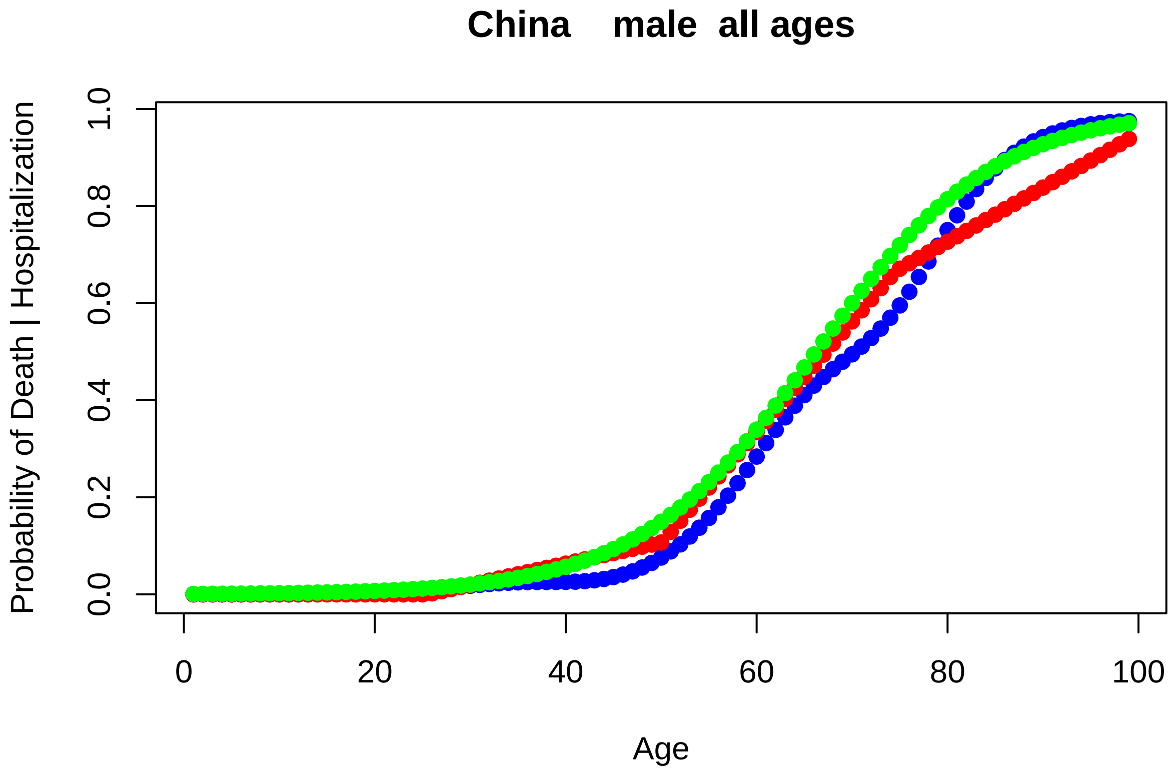
```
## [1] China
## 38 Levels: Afghanistan Algeria Australia Austria Bahrain Belgium ... Vietnam
```

```
CV_model[600, "country"]
```

```
## [1] Taiwan
## 38 Levels: Afghanistan Algeria Australia Austria Bahrain Belgium ... Vietnam
```

Back to our China specific probabilities for all ages…

```
plot(ages, np_predictions, pch = 19, col = 'blue', xlab = "Age",
     ylab = "Probability of Death | Hospitalization",
     main = paste0(country_of_interest,"  " ,"  ", gender_of_interest,"  ", "all ages"))
points(ages, NNS_predictions$Fitted.xy$y.hat, pch = 19, col = 'red')
points(ages, logit_predictions, pch = 19, col = 'green')
```



All 3 models give relatively similar predictions for all ages of Chinese males. Given that each of the models have a vastly different approach to conditional estimation, the robustness of these probabilities is pronounced.

Causal Analysis

Using `incubation_proxy` as a proxy variable for virus incubation, we perform a causal analysis using the methods from the `generalCorr` package in R. We create a new subset, eliminating `country` in order to use it as a control variable.

```
library(generalCorr)
CV_model_2 <- cbind(corona[, c("death", "country", "gender", "age")], incubation_proxy)
```

```
CV_model_2 <- CV_model_2[complete.cases(CV_model_2), ]


causeSummary(data.matrix(CV_model_2[, -2]),ctrl=as.numeric(CV_model_2$country))
```

```
## [1] gender    causes    death     strength= -37.008
## [1] corr=   0.1296  p-val=  0.29226
## [1] death     causes    age       strength= 31.496
## [1] corr=   -0.0441 p-val=  0.72129
## [1] incubation_proxy causes           death            strength=
## [5] -37.008
## [1] corr=   -0.0467 p-val=  0.70502
```

```
##      cause              response strength corr.     p-value
## [1,] "gender"           "death"  "37.008" "0.1296"  "0.29226"
## [2,] "death"            "age"    "31.496" "-0.0441" "0.72129"
## [3,] "incubation_proxy" "death"  "37.008" "-0.0467" "0.70502"
```

Unfortunately the causal evidence for the `incubation_proxy` contributing to `death` is not significant.

Comments

Obviously without more reliable / accurate data, these estimates are weak. But we do know the probabilities of death are indeed elevated for a hospitalization relating from a positive COVID-19 diagnosis. This is in contrast with the overall mortality rate being reported which includes those individuals who did not require hospitalization and this distinguishing condition should be duly noted.

The hospitalization this data reflects may indeed by an indication of an underlying pre-existing condition. This analysis supports the common sense advice that those individuals with such conditions be exceptionally prudent and precautionary in their navigation of these difficult times.