COVID-19 analysis based on crowdsourced data

Data source

Paper: Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study

Date downloaded: 2020-03-11. Latest data update: 2020-03-09 1PM EST.

DISCLAIMER: Data quality is **very questionable**. As such, any conclusions will be very questionable too. This is meant as a didactic exercise only. Students are invited to find better data and/or keep a very critical mindset while going through the results.

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------- tidy

## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0

## -- Conflicts ---------------------------------------------------------------------------- tidyverse_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data preparation

```
d_raw <- read_csv("covid19.csv",
              col_types = cols(id = 'c', case_in = 'c', age = 'd', if_onset_approximated = 'l',
                               international_traveler = 'l', domestic_traveler = 'l', traveler = 'l',
                               `visiting Wuhan` = 'l', `from Wuhan` = 'l',
                               .default = 'c'
                               ))
```

```
## Warning: 1 parsing failure.
##  row                   col           expected actual          file
## 2224 if_onset_approximated 1/0/T/F/TRUE/FALSE     53 'covid19.csv'
```

```
#?
#d_raw <- d_raw[-2224,]
```

```
table(d_raw$death)
```

```
##
##         0         1  2/1/2020 2/13/2020 2/14/2020 2/19/2020 2/21/2020
##      1536        42         1         1         1         2         2
## 2/22/2020 2/23/2020 2/24/2020 2/25/2020 2/26/2020 2/27/2020 2/28/2020
##         2         5         1         2         4         2         1
## 2/29/2020  3/1/2020  3/3/2020  3/4/2020  3/6/2020  3/8/2020
##         1         3         1         1         1         2
```

```
as_date <- function(x) as.Date(x, format = "%m/%d/%y")
d <-
  d_raw %>%
  mutate(reporting_date = as_date(reporting_date),
         hosp_visit_date = as_date(hosp_visit_date),
         exposure_start = as_date(exposure_start),
         exposure_end = as_date(exposure_end),
         symptom_onset = as_date(symptom_onset),
         death_status = death != "0",
         death_date = as.Date(ifelse(!death %in% c("0", "1"), as.Date(death, format = "%m/%d/%y", origi
         gender = factor(gender, levels = c("female", "male")))
```

Binary outcome: alive/dead

Sex impact

Frequency tables and independence tests

(1) Chi-Square: p-value > alpha = 0.05, cannot reject null hypothesis, gender and death are independent

```
with(d, table(death_status, gender))
```

```
##              gender
## death_status female male
##        FALSE    516  664
##        TRUE      22   47
```

```
with(d, prop.table(table(death_status, gender), 2))
```

```
##              gender
## death_status     female       male
##        FALSE 0.95910781 0.93389592
##        TRUE  0.04089219 0.06610408
```

```
with(d, 100 * prop.table(table(death_status, gender), 2))
```

```
##              gender
## death_status    female      male
##        FALSE 95.910781 93.389592
##        TRUE   4.089219  6.610408
```

```
with(d, round(100 * prop.table(table(death_status, gender), 2), 1))
```

```
##              gender
## death_status female male
##        FALSE   95.9 93.4
##        TRUE     4.1  6.6
```

```r
with(d, chisq.test(table(death_status, gender)))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(death_status, gender)
## X-squared = 3.2625, df = 1, p-value = 0.07088
```

(2) Fisher Test: p-value > alpha = 0.05, cannot reject null hypothesis, gender and death are independent

```r
with(d, fisher.test(table(death_status, gender)))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(death_status, gender)
## p-value = 0.06051
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9660476 2.9313166
## sample estimates:
## odds ratio
##   1.659538
```

Logistic regression

```r
summary(lm(death_status ~ gender, data = d))
```

```
##
## Call:
## lm(formula = death_status ~ gender, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06610 -0.06610 -0.06610 -0.04089  0.95911
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.040892   0.009843   4.155 3.48e-05 ***
## gendermale  0.025212   0.013045   1.933   0.0535 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2283 on 1247 degrees of freedom
##   (1061 observations deleted due to missingness)
## Multiple R-squared:  0.002986,   Adjusted R-squared:  0.002187
## F-statistic: 3.735 on 1 and 1247 DF,  p-value: 0.05351
```

```
summary(glm(death_status ~ gender, data = d, family = "binomial"))
```

```
##
## Call:
## glm(formula = death_status ~ gender, family = "binomial", data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3698  -0.3698  -0.3698  -0.2890   2.5286
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.1551     0.2176 -14.496   <2e-16 ***
## gendermale    0.5069     0.2649   1.914   0.0556 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 533.76  on 1248  degrees of freedom
## Residual deviance: 529.92  on 1247  degrees of freedom
##   (1061 observations deleted due to missingness)
## AIC: 533.92
##
## Number of Fisher Scoring iterations: 5
```

```
exp(confint(glm(death_status ~ gender, data = d, family = "binomial")))
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %    97.5 %
## (Intercept) 0.02700237 0.0636529
## gendermale  0.99988099 2.8395321
```

```
exp(confint(glm(death_status ~ gender, data = d, family = "binomial"))[2,])
```

```
## Waiting for profiling to be done...
```

```
##    2.5 %   97.5 %
## 0.999881 2.839532
```

```
confint(glm(death_status ~ gender, data = d, family = "binomial"))
```

```
## Waiting for profiling to be done...
```

```
##                     2.5 %    97.5 %
## (Intercept) -3.6118308097 -2.754310
## gendermale  -0.0001190204  1.043639
```

```r
exp(0.5069)
```

```
## [1] 1.660137
```

Age

```r
glm(death_status ~ age, data = d, family = "binomial")
```

```
## 
## Call:  glm(formula = death_status ~ age, family = "binomial", data = d)
## 
## Coefficients:
## (Intercept)          age  
##    -7.56441      0.07989  
## 
## Degrees of Freedom: 1159 Total (i.e. Null);  1158 Residual
##   (1150 observations deleted due to missingness)
## Null Deviance:       528.8 
## Residual Deviance: 427.2     AIC: 431.2
```

```r
summary(glm(death_status ~ age, data = d, family = "binomial"))
```

```
## 
## Call:
## glm(formula = death_status ~ age, family = "binomial", data = d)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.2225  -0.3623  -0.2094  -0.1143   3.0651  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -7.564414   0.652945 -11.585   <2e-16 ***
## age          0.079895   0.009362   8.534   <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: 528.76  on 1159  degrees of freedom
## Residual deviance: 427.16  on 1158  degrees of freedom
##   (1150 observations deleted due to missingness)
## AIC: 431.16
## 
## Number of Fisher Scoring iterations: 7
```

```r
fit <- glm(death_status ~ I(age/10), data = d, family = "binomial")
summary(fit)
```

```
## 
## Call:
## glm(formula = death_status ~ I(age/10), family = "binomial", 
##     data = d)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.2225  -0.3623  -0.2094  -0.1143   3.0651  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -7.56441    0.65295 -11.585   <2e-16 ***
## I(age/10)    0.79895    0.09362   8.534   <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: 528.76  on 1159  degrees of freedom
## Residual deviance: 427.16  on 1158  degrees of freedom
##   (1150 observations deleted due to missingness)
## AIC: 431.16
## 
## Number of Fisher Scoring iterations: 7
```

```
exp(confint(fit)[2,])
```

```
## Waiting for profiling to be done...
```

```
##    2.5 %   97.5 % 
## 1.863645 2.691992
```

```
exp(coef(fit)[2])
```

```
## I(age/10) 
##  2.223203
```

Smoothing splines for age

```
library(mgcv)
```

```
## Loading required package: nlme
```

```
## 
## Attaching package: 'nlme'
```

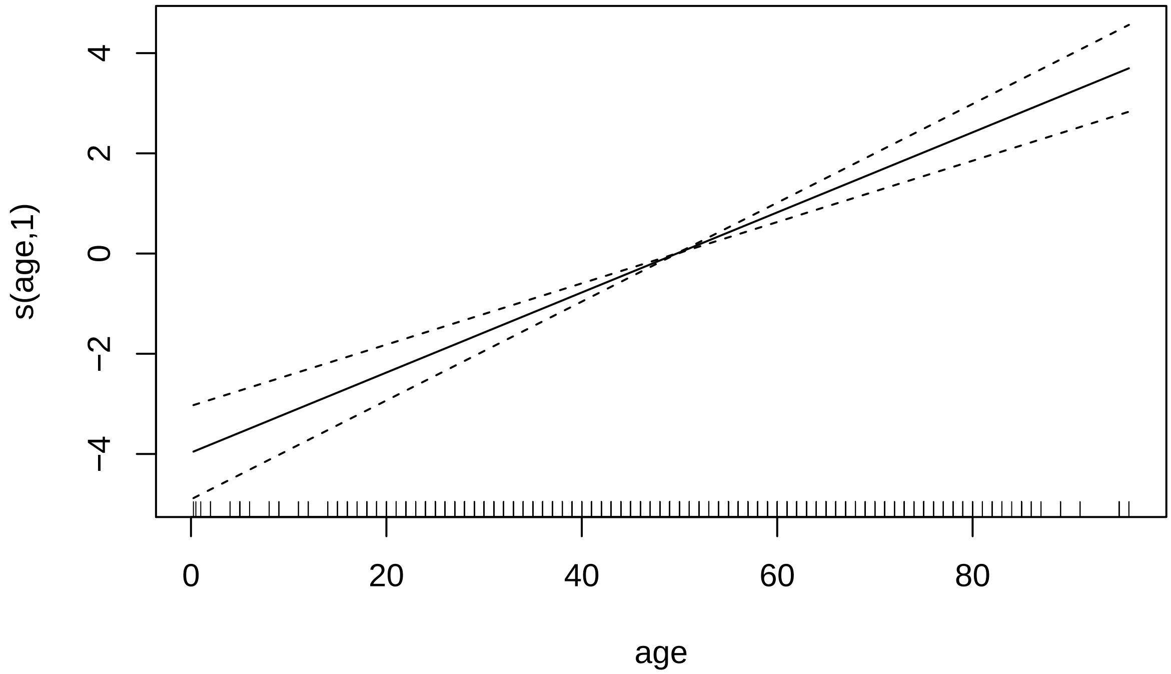
```
## The following object is masked from 'package:dplyr':
## 
##     collapse
## This is mgcv 1.8-26. For overview type 'help("mgcv-package")'.
```

```
fit <- gam(death_status ~ s(age), data = d, family = "binomial")
```

```
summary(fit)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## death_status ~ s(age)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.592      0.218  -16.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##        edf Ref.df Chi.sq p-value
## s(age)   1  1.001  72.79  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.108   Deviance explained = 19.2%
## UBRE = -0.62831  Scale est. = 1         n = 1160
```

```
plot(fit)
```



Quadratic term for age

```
fit <- glm(death_status ~ age + I(age^2), data = d, family = "binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = death_status ~ age + I(age^2), family = "binomial",
##     data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1334  -0.3713  -0.2060  -0.1010   3.1212
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.6688348  2.1839829  -3.969 7.21e-05 ***
## age          0.1156056  0.0671105   1.723    0.085 .
## I(age^2)    -0.0002749  0.0005074  -0.542    0.588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 528.76  on 1159  degrees of freedom
## Residual deviance: 426.85  on 1157  degrees of freedom
##   (1150 observations deleted due to missingness)
## AIC: 432.85
##
## Number of Fisher Scoring iterations: 8
```

Piece-wise linear terms

```
d2 <- mutate(d, age_70 = ifelse(age > 70, age - 70, 0))
```

```
fit <- glm(death_status ~ age + age_70, data = d2, family = "binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = death_status ~ age + age_70, family = "binomial",
##     data = d2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1294  -0.3689  -0.2018  -0.1037   3.1170
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.01321    1.01214  -7.917 2.43e-15 ***
## age          0.08787    0.01641   5.354 8.58e-08 ***
## age_70      -0.02061    0.03363  -0.613     0.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 528.76  on 1159  degrees of freedom
## Residual deviance: 426.78  on 1157  degrees of freedom
##   (1150 observations deleted due to missingness)
## AIC: 432.78
##
## Number of Fisher Scoring iterations: 7
```

From 0 to 70:

```
exp(coef(fit)[2])
```

```
##      age
## 1.091842
```

After 70:

```r
exp(coef(fit)[2:3])
```

```
##       age    age_70 
## 1.0918416 0.9796014
```

```r
exp(sum(coef(fit)[2:3]))
```

```
## [1] 1.06957
```

A different parametrization

```r
d3 <- mutate(d,
             age_l70 = ifelse(age <= 70, age, 70),
             age_g70 = ifelse(age > 70, age - 70, 0))
fit <- glm(death_status ~ age_l70 + age_g70, data = d3, family = "binomial")

summary(fit)
```

```
## 
## Call:
## glm(formula = death_status ~ age_l70 + age_g70, family = "binomial", 
##     data = d3)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.1294  -0.3689  -0.2018  -0.1037   3.1170  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -8.01321    1.01214  -7.917 2.43e-15 ***
## age_l70      0.08787    0.01641   5.354 8.58e-08 ***
## age_g70      0.06726    0.02258   2.979   0.0029 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 528.76  on 1159  degrees of freedom
## Residual deviance: 426.78  on 1157  degrees of freedom
##   (1150 observations deleted due to missingness)
## AIC: 432.78
## 
## Number of Fisher Scoring iterations: 7
```

```r
exp(confint(fit)[2:3,])
```

```
## Waiting for profiling to be done...
```

```
##            2.5 %   97.5 %
## age_l70 1.059851 1.130634
## age_g70 1.022748 1.117977
```

Comparing countries

Caveats: 1. obsolete data 2. different testing guidelines per country 3. different reporting accuracy (i.e., often only deaths from hospitals are counted) 4. . . . ?

```
table(d$country)
```

```
##
## Afghanistan     Algeria   Australia     Austria     Bahrain     Belgium
##           1           1          15           2          17           1
##    Cambodia      Canada       China     Croatia       Egypt     Finland
##           1          12         197           1           1           1
##      France     Germany   Hong Kong       India        Iran      Israel
##          56         168         102           3          18           1
##       Italy       Japan      Kuwait     Lebanon    Malaysia       Nepal
##          86         257           9           1          23           1
## Phillipines      Russia   Singapore South Korea       Spain   Sri Lanka
##           3           2         112         114         116           1
##      Sweden Switzerland      Taiwan    Thailand         UAE          UK
##           1          10          34          41          21          20
##         USA     Vietnam
##         757          16
```

```
d1 <- mutate(d, country = relevel(factor(country), ref = "China"))
summary(glm(death_status ~ country, data = d1, family = "binomial"))
```

```
##
## Call:
## glm(formula = death_status ~ country, family = "binomial", data = d1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.90052  -0.37804  -0.19823  -0.00005   2.80700
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.3990     0.1788  -7.824  5.1e-15 ***
## countryAfghanistan   -19.1670 17730.3699  -0.001 0.999137
## countryAlgeria       -19.1670 17730.3699  -0.001 0.999137
## countryAustralia     -19.1670  4577.9618  -0.004 0.996659
## countryAustria       -19.1670 12537.2648  -0.002 0.998780
## countryBahrain       -19.1670  4300.2464  -0.004 0.996444
## countryBelgium       -19.1670 17730.3699  -0.001 0.999137
## countryCambodia      -19.1670 17730.3699  -0.001 0.999137
## countryCanada        -19.1670  5118.3169  -0.004 0.997012
## countryCroatia       -19.1670 17730.3699  -0.001 0.999137
## countryEgypt         -19.1670 17730.3699  -0.001 0.999137
## countryFinland       -19.1670 17730.3699  -0.001 0.999137
## countryFrance         -1.8968     0.7419  -2.557 0.010573 *
## countryGermany       -19.1670  1367.9277  -0.014 0.988821
## countryHong Kong      -2.5130     0.7362  -3.414 0.000641 ***
## countryIndia         -19.1670 10236.6338  -0.002 0.998506
## countryIran            0.1463     0.5945   0.246 0.805643
## countryIsrael        -19.1670 17730.3699  -0.001 0.999137    
## countryItaly          -2.3386     0.7375  -3.171 0.001519 ** 
## countryJapan          -2.5210     0.4857  -5.190  2.1e-07 ***
## countryKuwait        -19.1670  5910.1233  -0.003 0.997412    
## countryLebanon       -19.1670 17730.3699  -0.001 0.999137    
## countryMalaysia      -19.1670  3697.0377  -0.005 0.995863    
## countryNepal         -19.1670 17730.3699  -0.001 0.999137    
## countryPhillipines     0.7059     1.2377   0.570 0.568469    
## countryRussia        -19.1670 12537.2648  -0.002 0.998780    
## countrySingapore     -19.1670  1675.3625  -0.011 0.990872    
## countrySouth Korea    -1.0577     0.3906  -2.708 0.006778 ** 
## countrySpain         -19.1670  1646.2235  -0.012 0.990710    
## countrySri Lanka     -19.1670 17730.3699  -0.001 0.999137    
## countrySweden        -19.1670 17730.3699  -0.001 0.999137    
## countrySwitzerland   -19.1670  5606.8353  -0.003 0.997272    
## countryTaiwan         -2.0975     1.0307  -2.035 0.041844 *  
## countryThailand      -19.1670  2769.0186  -0.007 0.994477    
## countryUAE           -19.1670  3869.0839  -0.005 0.996047    
## countryUK            -19.1670  3964.6312  -0.005 0.996143    
## countryUSA            -1.2037     0.3733  -3.224 0.001264 ** 
## countryVietnam       -19.1670  4432.5925  -0.004 0.996550    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 606.52  on 1610  degrees of freedom
## Residual deviance: 468.95  on 1573  degrees of freedom
##   (699 observations deleted due to missingness)
## AIC: 544.95
## 
## Number of Fisher Scoring iterations: 19
```

Adjusted model

```
fit <- glm(death_status ~ I(age/10) + gender + country, data = d1, family = "binomial")
summary(fit)
```

```
## 
## Call:
## glm(formula = death_status ~ I(age/10) + gender + country, family = "binomial", 
##     data = d1)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.6949  -0.1932  -0.0610   0.0000   3.6227  
## 
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)    
## (Intercept)              -7.9714     0.9070  -8.789  < 2e-16 ***
## I(age/10)                 1.0603     0.1303   8.135 4.11e-16 ***
## gendermale                0.8658     0.3441   2.516  0.01186 *  
## countryAustralia     -18.4937  4123.7162  -0.004  0.99642
## countryCambodia      -19.8222 17730.3699  -0.001  0.99911
## countryCanada        -18.0770  4730.8786  -0.004  0.99695
## countryFinland       -15.9876 17730.3699  -0.001  0.99928
## countryFrance         -0.7951     0.9138  -0.870  0.38425
## countryGermany       -19.2704  1550.0236  -0.012  0.99008
## countryHong Kong      -3.5901     0.7891  -4.550 5.37e-06 ***
## countryItaly          -2.1922     0.8409  -2.607  0.00913 **
## countryJapan          -3.9523     0.5970  -6.620 3.59e-11 ***
## countryLebanon       -17.3660 17730.3699  -0.001  0.99922
## countryMalaysia      -18.7206  3172.3686  -0.006  0.99529
## countryNepal         -16.8534 17730.3699  -0.001  0.99924
## countryPhillipines     1.8041     1.3598   1.327  0.18461
## countrySingapore     -18.5859  1527.9201  -0.012  0.99029
## countrySouth Korea    -0.7657     0.4593  -1.667  0.09551 .
## countrySpain         -19.8696  1761.9807  -0.011  0.99100
## countrySri Lanka     -16.8359 17730.3699  -0.001  0.99924
## countrySweden        -15.2454 17730.3699  -0.001  0.99931
## countrySwitzerland   -18.8654  6051.8537  -0.003  0.99751
## countryTaiwan         -2.5370     1.1298  -2.245  0.02474 *
## countryThailand      -19.3291  3947.2803  -0.005  0.99609
## countryUAE           -19.3137  5885.6028  -0.003  0.99738
## countryUK            -19.0800 17730.3699  -0.001  0.99914
## countryUSA            -1.0312     0.5283  -1.952  0.05092 .
## countryVietnam       -18.4301  5439.6772  -0.003  0.99730
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 518.77  on 1123  degrees of freedom
## Residual deviance: 278.36  on 1096  degrees of freedom
##   (1186 observations deleted due to missingness)
## AIC: 334.36
##
## Number of Fisher Scoring iterations: 19
```

```r
exp(coef(fit)[2:3])
```

```
##  I(age/10) gendermale
##   2.887235   2.376838
```

Maximum Likelihood Estimation

```r
library(maxLik)
```

```
## Loading required package: miscTools
```

```
##
## Please cite the 'maxLik' package as:
```

```
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. C
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum o
## https://r-forge.r-project.org/projects/maxlik/
```

Model: death by gender

```
dx <- d %>% select(gender, death_status) %>% na.omit()

x <- (dx$gender == "male") + 0
y <- dx$death_status + 0

logLik <- function(beta) {
  linear_predictor <- beta[1] + beta[2] * x
  log_probabilities <-
    dbinom(y, size = 1, prob = plogis(linear_predictor), log = TRUE)
  log_likelihood <- sum(log_probabilities)
  return(log_likelihood)
}
```

```
debugonce(logLik)
```

Check for the accuracy:

```
logLik(c(0, 0))
```

```
## debugging in: logLik(c(0, 0))
## debug at <text>#6: {
##     linear_predictor <- beta[1] + beta[2] * x
##     log_probabilities <- dbinom(y, size = 1, prob = plogis(linear_predictor),
##         log = TRUE)
##     log_likelihood <- sum(log_probabilities)
##     return(log_likelihood)
## }
## debug at <text>#7: linear_predictor <- beta[1] + beta[2] * x
## debug at <text>#8: log_probabilities <- dbinom(y, size = 1, prob = plogis(linear_predictor),
##     log = TRUE)
## debug at <text>#10: log_likelihood <- sum(log_probabilities)
## debug at <text>#11: return(log_likelihood)
## exiting from: logLik(c(0, 0))
```

```
## [1] -865.7408
```

```
fit <- maxLik(logLik, start = c(intercept = 0, gender = 0))
```

```
summary(fit)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 6 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -264.9617
## 2  free parameters
## Estimates:
##           Estimate Std. error t value Pr(> t)
## intercept  -3.1551     0.2181 -14.469  <2e-16 ***
## gender      0.5069     0.2655   1.909  0.0562 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

```
summary(glm(death_status ~ gender, data = dx, family = "binomial"))
```

```
##
## Call:
## glm(formula = death_status ~ gender, family = "binomial", data = dx)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3698  -0.3698  -0.3698  -0.2890   2.5286
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.1551     0.2176 -14.496   <2e-16 ***
## gendermale    0.5069     0.2649   1.914   0.0556 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 533.76  on 1248  degrees of freedom
## Residual deviance: 529.92  on 1247  degrees of freedom
## AIC: 533.92
##
## Number of Fisher Scoring iterations: 5
```

Continuous outcome: survival time

```
library(survival)
```

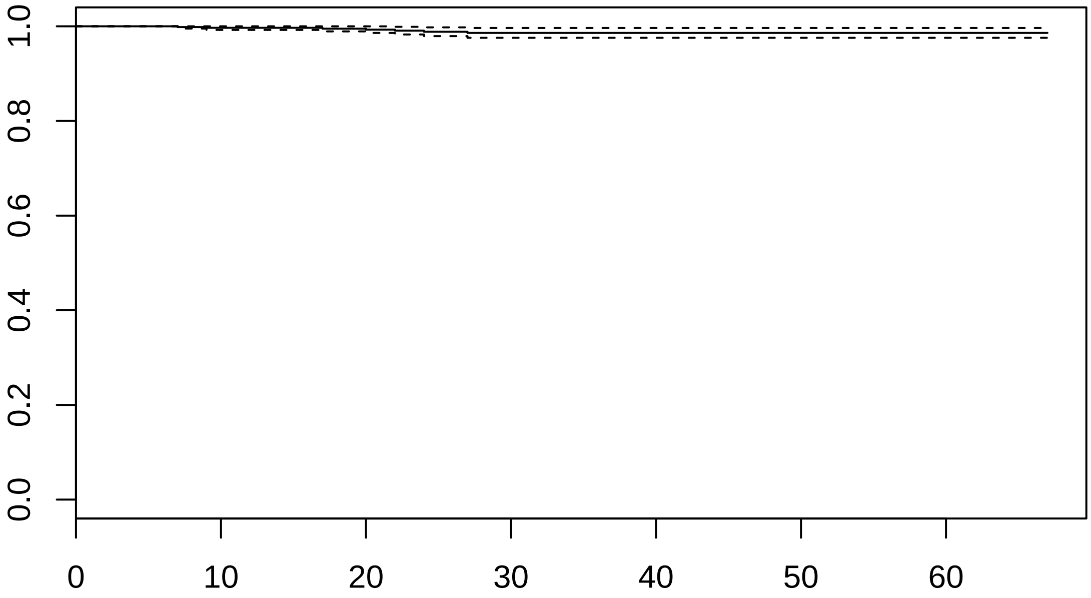
```
##
## Attaching package: 'survival'

## The following object is masked from 'package:rpart':
##
##     solder
```

```
END_OF_STUDY <- as.Date("2020-03-10", format = "%Y-%m-%d", origin = "1970-01-01")
```

From first symptom to date of data collection: March 10th, 2020

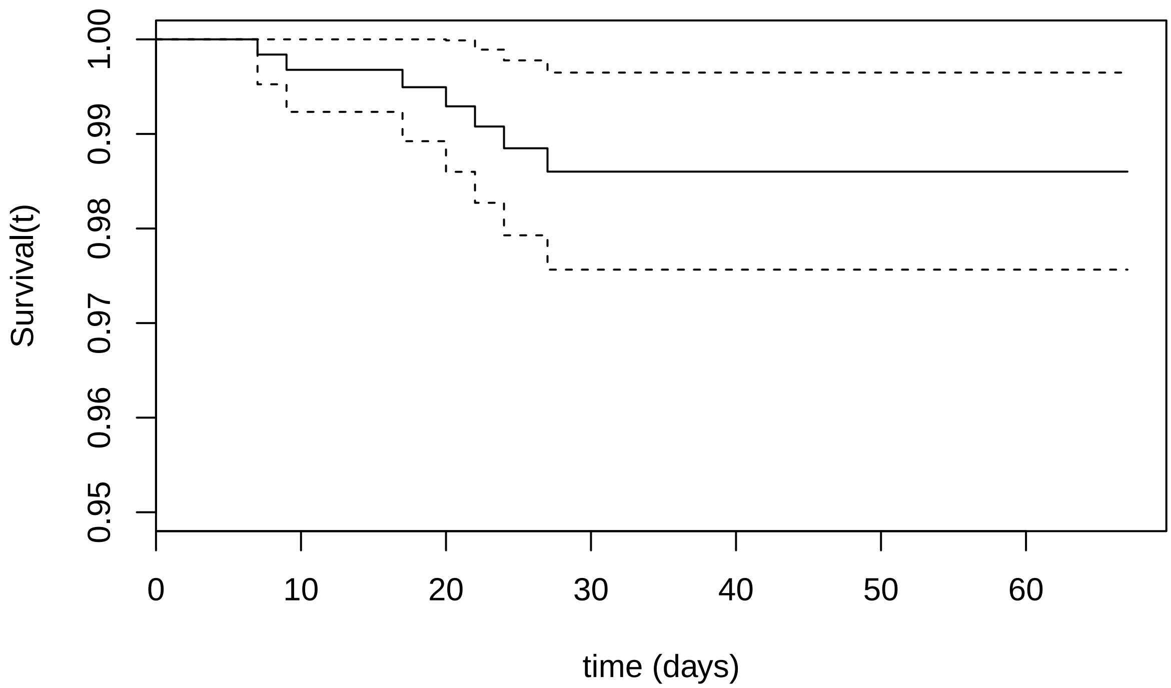
```
d_surv <-
  d %>%
  filter(!is.na(symptom_onset), !is.na(death_status),
         !(is.na(death_date) & death_status)) %>%
  mutate(
    death_date = as.Date(ifelse(is.na(death_date), END_OF_STUDY, death_date), origin = "1970-01-01"),
    time = difftime(death_date, symptom_onset, units = "days")
  )
```

```
fit <- survfit(Surv(time, death_status) ~ 1, data = d_surv)
plot(fit)
```



Let’s zoom in:

```
plot(fit, ylim = c(0.95, 1),
     xlab = "time (days)", ylab = "Survival(t)")
```



Cannot extract a median, 7 deaths out of 623 cases (~1.1%):

```
fit
```

```
## Call: survfit(formula = Surv(time, death_status) ~ 1, data = d_surv)
##
##       n  events  median 0.95LCL 0.95UCL
##     623       7      NA      NA      NA
```

Effects of covariates on risk of death

```
summary(d_surv)
```

```
##       id              case_in          reporting_date
##  Length:623         Length:623         Min.   :2020-01-13
##  Class :character   Class :character   1st Qu.:2020-01-27
##  Mode  :character   Mode  :character   Median :2020-02-14
##                                        Mean   :2020-02-11
##                                        3rd Qu.:2020-02-25
##                                        Max.   :2020-03-06
##
##    summary            location           country             gender
##  Length:623         Length:623         Length:623         female:252
##  Class :character   Class :character   Class :character   male  :365
##  Mode  :character   Mode  :character   Mode  :character   NA's  :  6
##
##
##
##
##       age       symptom_onset        if_onset_approximated
##  Min.   : 2.0   Min.   :2020-01-03   Mode :logical
##  1st Qu.:37.0   1st Qu.:2020-01-23   FALSE:574
##  Median :52.0   Median :2020-02-03   TRUE :24
##  Mean   :50.3   Mean   :2020-02-04   NA's :25
##  3rd Qu.:65.0   3rd Qu.:2020-02-17
##  Max.   :96.0   Max.   :2020-03-05
##  NA's   :18
##  hosp_visit_date      exposure_start        exposure_end
##  Min.   :2020-01-06   Min.   :2020-01-03   Min.   :2020-01-02
##  1st Qu.:2020-01-25   1st Qu.:2020-01-12   1st Qu.:2020-01-17
##  Median :2020-02-06   Median :2020-01-20   Median :2020-01-21
##  Mean   :2020-02-07   Mean   :2020-02-02   Mean   :2020-01-22
##  3rd Qu.:2020-02-20   3rd Qu.:2020-01-25   3rd Qu.:2020-01-23
##  Max.   :2020-03-02   Max.   :2020-12-29   Max.   :2020-03-04
##  NA's   :131          NA's   :536          NA's   :407
##  international_traveler domestic_traveler  traveler       visiting Wuhan
##  Mode :logical          Mode :logical     Mode :logical   Mode :logical
##  FALSE:2                FALSE:8           FALSE:57        FALSE:507
##  TRUE :5                NA's :615         TRUE :118       TRUE :116
##  NA's :616                                NA's :448
##
##
##
##  from Wuhan         death           recovered          symptom
##  Mode :logical   Length:623         Length:623         Length:623
##  FALSE:543       Class :character   Class :character   Class :character
##  TRUE :76        Mode  :character   Mode  :character   Mode  :character
##  NA's :4
##
##
##
##     source              link           death_status
##  Length:623         Length:623         Mode :logical
##  Class :character   Class :character   FALSE:616
##  Mode  :character   Mode  :character   TRUE :7
##
##
##
##
##    death_date             time
##  Min.   :2020-02-01   Length:623
##  1st Qu.:2020-03-10   Class :difftime
##  Median :2020-03-10   Mode  :numeric
##  Mean   :2020-03-09
##  3rd Qu.:2020-03-10
##  Max.   :2020-03-10
##
```

Gender

```r
summary(coxph(Surv(time, death_status) ~ gender, data = d_surv))
```

```
## Call:
## coxph(formula = Surv(time, death_status) ~ gender, data = d_surv)
##
##   n= 617, number of events= 6
##    (6 observations deleted due to missingness)
##
##              coef exp(coef) se(coef)     z Pr(>|z|)
## gendermale 1.256     3.513    1.095 1.147    0.251
##
##            exp(coef) exp(-coef) lower .95 upper .95
## gendermale     3.513     0.2847    0.4104     30.07
##
## Concordance= 0.628  (se = 0.073 )
## Rsquare= 0.003   (max possible= 0.113 )
## Likelihood ratio test= 1.68  on 1 df,   p=0.2
## Wald test            = 1.32  on 1 df,   p=0.3
## Score (logrank) test = 1.5  on 1 df,   p=0.2
```

Country

Model does not converge: again, too few cases.

```r
summary(coxph(Surv(time, death_status) ~ country, data = d_surv))
```

```
## Warning in fitter(X, Y, strats, offset, init, control, weights = weights, :
## Ran out of iterations and did not converge
```

```
## Call:
## coxph(formula = Surv(time, death_status) ~ country, data = d_surv)
##
##   n= 623, number of events= 7
##
##                         coef exp(coef)  se(coef)     z Pr(>|z|)
## countryCambodia    1.435e-06 1.000e+00 1.306e+05 0.000        1
## countryCanada      1.649e-06 1.000e+00 9.685e+04 0.000        1
## countryChina       1.614e-06 1.000e+00 4.294e+04 0.000        1
## countryFinland     1.450e-06 1.000e+00 1.306e+05 0.000        1
## countryFrance      2.222e+01 4.471e+09 4.130e+04 0.001        1
## countryGermany     1.138e-01 1.120e+00 6.603e+04 0.000        1
## countryHong Kong   2.874e-02 1.029e+00 4.389e+04 0.000        1
## countryItaly       9.368e-02 1.098e+00 1.443e+05 0.000        1
## countryJapan       2.034e+01 6.841e+08 4.130e+04 0.000        1
## countryMalaysia    1.555e-02 1.016e+00 5.747e+04 0.000        1
## countryNepal       1.442e-06 1.000e+00 1.306e+05 0.000        1
## countryPhillipines 1.558e+02 4.575e+67 0.000e+00   Inf   <2e-16 ***
## countrySingapore   2.525e-02 1.026e+00 4.348e+04 0.000        1
## countrySouth Korea 1.455e-06 1.000e+00 5.107e+04 0.000        1
## countrySpain       4.908e-01 1.634e+00 3.292e+05 0.000        1
## countrySri Lanka   1.429e-06 1.000e+00 1.306e+05 0.000        1
## countrySweden      1.440e-06 1.000e+00 1.306e+05 0.000        1
## countrySwitzerland 1.817e-01 1.199e+00 1.185e+05 0.000        1
## countryTaiwan      3.428e-03 1.003e+00 4.897e+04 0.000        1
## countryThailand    1.604e-06 1.000e+00 6.244e+04 0.000        1
## countryUAE         1.374e-06 1.000e+00 1.306e+05 0.000        1
## countryUSA         1.059e-02 1.011e+00 6.203e+04 0.000        1
## countryVietnam     1.512e-06 1.000e+00 6.910e+04 0.000        1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## countryCambodia    1.000e+00  1.000e+00 0.000e+00       Inf
## countryCanada      1.000e+00  1.000e+00 0.000e+00       Inf
## countryChina       1.000e+00  1.000e+00 0.000e+00       Inf
## countryFinland     1.000e+00  1.000e+00 0.000e+00       Inf
## countryFrance      4.471e+09  2.236e-10 0.000e+00       Inf
## countryGermany     1.120e+00  8.925e-01 0.000e+00       Inf
## countryHong Kong   1.029e+00  9.717e-01 0.000e+00       Inf
## countryItaly       1.098e+00  9.106e-01 0.000e+00       Inf
## countryJapan       6.841e+08  1.462e-09 0.000e+00       Inf
## countryMalaysia    1.016e+00  9.846e-01 0.000e+00       Inf
## countryNepal       1.000e+00  1.000e+00 0.000e+00       Inf
## countryPhillipines 4.575e+67  2.186e-68 4.575e+67 4.575e+67
## countrySingapore   1.026e+00  9.751e-01 0.000e+00       Inf
## countrySouth Korea 1.000e+00  1.000e+00 0.000e+00       Inf
## countrySpain       1.634e+00  6.121e-01 0.000e+00       Inf
## countrySri Lanka   1.000e+00  1.000e+00 0.000e+00       Inf
## countrySweden      1.000e+00  1.000e+00 0.000e+00       Inf
## countrySwitzerland 1.199e+00  8.339e-01 0.000e+00       Inf
## countryTaiwan      1.003e+00  9.966e-01 0.000e+00       Inf
## countryThailand    1.000e+00  1.000e+00 0.000e+00       Inf
## countryUAE         1.000e+00  1.000e+00 0.000e+00       Inf
## countryUSA         1.011e+00  9.895e-01 0.000e+00       Inf
## countryVietnam     1.000e+00  1.000e+00 0.000e+00       Inf
##
## Concordance= 0.904  (se = 0.029 )
## Rsquare= 0.049   (max possible= 0.13 )
## Likelihood ratio test= 31.22  on 23 df,   p=0.1
## Wald test            = 2.91  on 23 df,   p=1
## Score (logrank) test = 648.9  on 23 df,   p=<2e-16
```

Is risk in France really 1.8 **billion** times that in China?

Age

This one is strong, as expected:

```
summary(coxph(Surv(time, death_status) ~ I(age / 10), data = d_surv))
```

```
## Call:
```

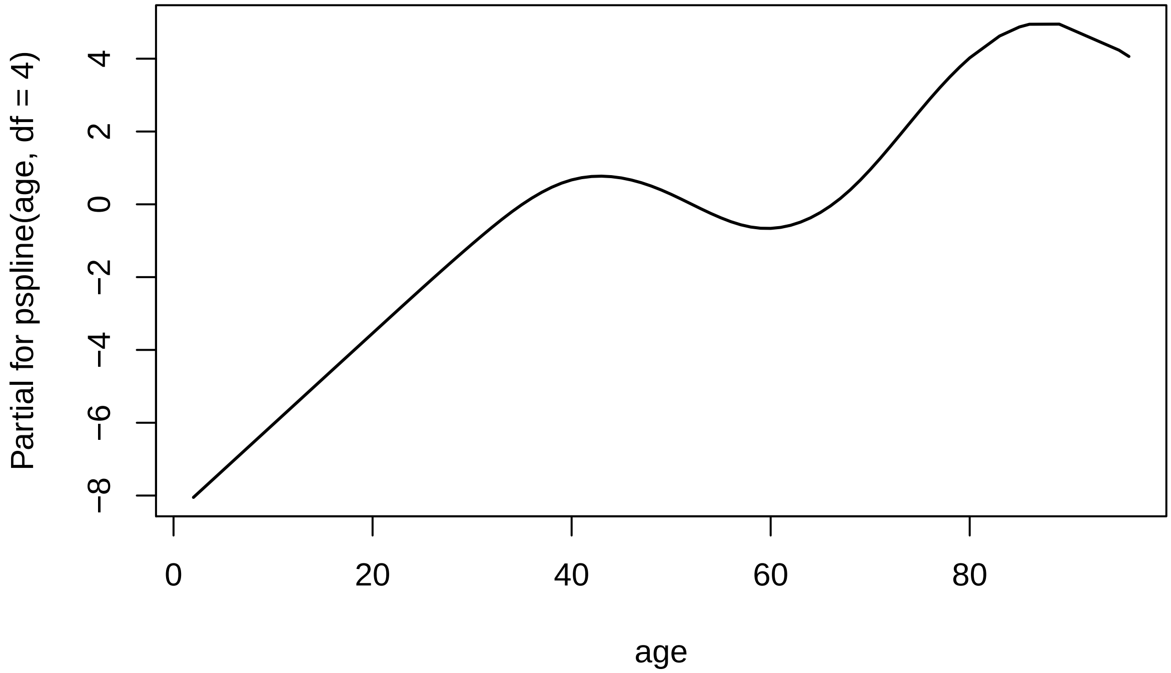
```
## coxph(formula = Surv(time, death_status) ~ I(age/10), data = d_surv)
## 
##   n= 605, number of events= 6 
##    (18 observations deleted due to missingness)
## 
##              coef exp(coef) se(coef)     z Pr(>|z|)    
## I(age/10) 1.2403    3.4567   0.3435 3.611 0.000305 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##           exp(coef) exp(-coef) lower .95 upper .95
## I(age/10)     3.457     0.2893     1.763     6.777
## 
## Concordance= 0.846  (se = 0.117 )
## Rsquare= 0.028   (max possible= 0.115 )
## Likelihood ratio test= 17.18  on 1 df,   p=3e-05
## Wald test            = 13.04  on 1 df,   p=3e-04
## Score (logrank) test = 14.43  on 1 df,   p=1e-04
```

Is the effect linear, though?

```
fit_age <- coxph(Surv(time, death_status) ~ pspline(age, df = 4), data = d_surv)
summary(fit_age)
```

```
## Call:
## coxph(formula = Surv(time, death_status) ~ pspline(age, df = 4), 
##     data = d_surv)
## 
##   n= 605, number of events= 6 
##    (18 observations deleted due to missingness)
## 
##                           coef  se(coef) se2     Chisq DF   p      
## pspline(age, df = 4), lin 0.108 0.02791  0.02791 14.97 1.00 0.00011
## pspline(age, df = 4), non                         5.43 2.96 0.14000
## 
##           exp(coef) exp(-coef) lower .95 upper .95
## ps(age)3  1.054e+01  9.488e-02 7.311e-08 1.519e+09
## ps(age)4  1.111e+02  9.003e-03 1.981e-13 6.227e+16
## ps(age)5  1.170e+03  8.545e-04 2.345e-17 5.839e+22
## ps(age)6  1.220e+04  8.197e-05 2.450e-19 6.073e+26
## ps(age)7  9.509e+04  1.052e-05 1.929e-19 4.688e+28
## ps(age)8  6.028e+04  1.659e-05 1.395e-19 2.604e+28
## ps(age)9  1.111e+04  9.002e-05 3.372e-20 3.660e+27
## ps(age)10 3.023e+04  3.308e-05 1.017e-19 8.980e+27
## ps(age)11 1.133e+06  8.824e-07 3.918e-18 3.278e+29
## ps(age)12 8.386e+06  1.193e-07 2.915e-17 2.413e+30
## ps(age)13 2.107e+06  4.747e-07 6.324e-18 7.017e+29
## ps(age)14 3.024e+05  3.306e-06 3.371e-19 2.713e+29
## 
## Iterations: 4 outer, 18 Newton-Raphson
##      Theta= 0.03932137 
## Degrees of freedom for terms= 4 
## Concordance= 0.95  (se = 0.95 )
## Likelihood ratio test= 25.26  on 3.96 df,   p=4e-05
```

```
termplot(fit_age, col.term = 1, col.se = 1)
```



Piecewise-linear age effect

Segments: constant 0-70; increasing after 70

```
e <-
  d_surv %>%
  mutate(
    age_70p = ifelse(age <= 70, 0, age - 70)
  )
```

```
fit_age_segments <- coxph(Surv(time, death_status) ~ age_70p, data = e)
summary(fit_age_segments)
```

```
## Call:
## coxph(formula = Surv(time, death_status) ~ age_70p, data = e)
##
##   n= 605, number of events= 6
##    (18 observations deleted due to missingness)
##
##            coef exp(coef) se(coef)    z Pr(>|z|)
## age_70p 0.21373   1.23828  0.04063 5.26 1.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## age_70p     1.238     0.8076     1.144     1.341
##
## Concordance= 0.866  (se = 0.1 )
## Rsquare= 0.027   (max possible= 0.115 )
## Likelihood ratio test= 16.43  on 1 df,   p=5e-05
## Wald test            = 27.67  on 1 df,   p=1e-07
## Score (logrank) test = 66.68  on 1 df,   p=3e-16
```

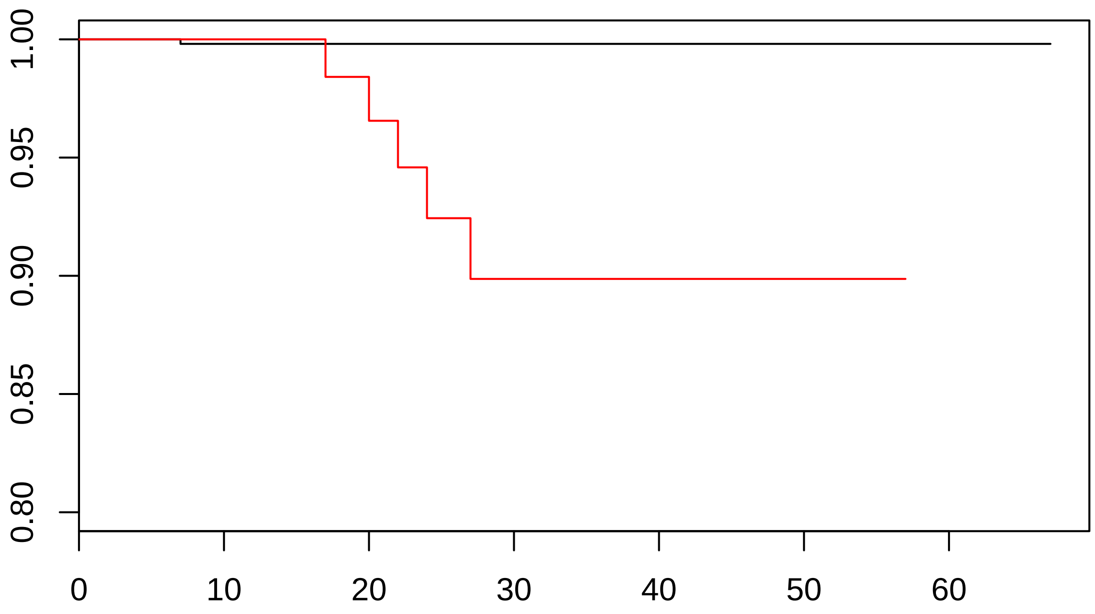
Or more simply, before and after 70:

```
fit_age_binary <- coxph(Surv(time, death_status) ~ I(age > 70), data = e)
summary(fit_age_binary)
```

```
## Call:
## coxph(formula = Surv(time, death_status) ~ I(age > 70), data = e)
##
##   n= 605, number of events= 6
##    (18 observations deleted due to missingness)
##
##                   coef exp(coef) se(coef)     z Pr(>|z|)
## I(age > 70)TRUE  3.694    40.200    1.097 3.368 0.000756 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## I(age > 70)TRUE      40.2    0.02488     4.686     344.9
##
## Concordance= 0.839  (se = 0.093 )
## Rsquare= 0.027   (max possible= 0.115 )
## Likelihood ratio test= 16.79  on 1 df,   p=4e-05
## Wald test            = 11.35  on 1 df,   p=8e-04
## Score (logrank) test = 31.52  on 1 df,   p=2e-08
```

Risk after 70yo is **40 times** that of people less than 70yo! Let's see it visually:

```
e$x <- factor(ifelse(e$age > 70, ">70", "<=70"), levels = c("<=70", ">70"))
plot(survfit(Surv(time, death_status) ~ I(age > 70), data = e),
     ylim = c(0.8, 1),
     col = 1:2)
```



```r
dx <- d%>% select(gender, death_status) %>% na.omit()

x <- (dx$gender == "male") +0
y <- dx$death_status +0

logLik <- function(beta) {
  linear_predictor = beta[1] +beta[2]

}
```