M1R example script: Regression analysis of COVID-19 data

Takoua Jendoubi

22 May 2020

Individual data from Kaggle

```r
indiv_data = read.csv(file="COVID19_open_line_list.csv")
```

We will first analyse sex vs the duration of disease onset to hospital admission then age vs the duration of disease onset to hospital admission Dates should be represented as an object of class Date in R. Inconsistent (negative) and missing dates are encoded NA by as.Date

```r
levels(indiv_data$date_onset_symptoms)
```

```
##  [1] ""                        "-25.02.2020"
##  [3] "01.01.2020"              "01.02.2020"
##  [5] "01.31.2020"              "02.01.2020"
##  [7] "02.02.2020"              "03.01.2020"
##  [9] "03.02.2020"              "04.01.2020"
## [11] "04.02.2020"              "04.04.2020"
## [13] "05.01.2020"              "05.02.2020"
## [15] "06.02.2020"              "07.02.2020"
## [17] "08.01.2020"              "08.02.2020"
## [19] "09.01.2020"              "09.02.2020"
## [21] "10.01.2020"              "10.01.2020 - 22.01.2020"
## [23] "10.02.2020"              "11.01.2020"
## [25] "11.02.2020"              "12.01.2020"
## [27] "12.02.2020"              "13.01.2020"
## [29] "13.02.2020"              "14.01.2020"
## [31] "14.02.2020"              "15.01.2020"
## [33] "15.02.2020"              "16.01.2020"
## [35] "16.02.2020"              "17.01.2020"
## [37] "17.02.2020"              "18.01.2020"
## [39] "18.02.2020"              "19.01.2020"
## [41] "19.02.2020"              "20.01.2020"
## [43] "20.02.2020"              "20.02.220"
## [45] "21.01.2020"              "21.02.2020"
## [47] "22.01.2020"              "22.02.2020"
## [49] "23.01.2020"              "23.02.2020"
## [51] "24.01.2020"              "24.02.2020"
## [53] "25.01.2020"              "25.02.2020"
## [55] "26.01.2020"              "26.02.2020"
## [57] "27.01.2020"              "27.02.2020"
## [59] "28.01.2020"              "29.01.2020"
## [61] "29.12.2019"              "30.01.2020"
## [63] "31.01.2020"              "end of December 2019"
## [65] "N/A"                     "none"
```

```
date_cols =  grep("date", colnames(indiv_data))
indiv_data[, date_cols] = lapply(indiv_data[, date_cols], as.Date, format="%d.%m.%y")
```

Define a new variable: the duration of disease onset to hospital admission

```
indiv_data$onset_to_hospital_ad = as.numeric(difftime(indiv_data$date_admission_hospital,
                                                      indiv_data$date_onset_symptoms, units="days"))
# Some durations are negative! Remove
indiv_data = indiv_data[indiv_data$onset_to_hospital_ad>=0,]
```

Data is inconsistent

```
table(as.factor(indiv_data$sex))
```

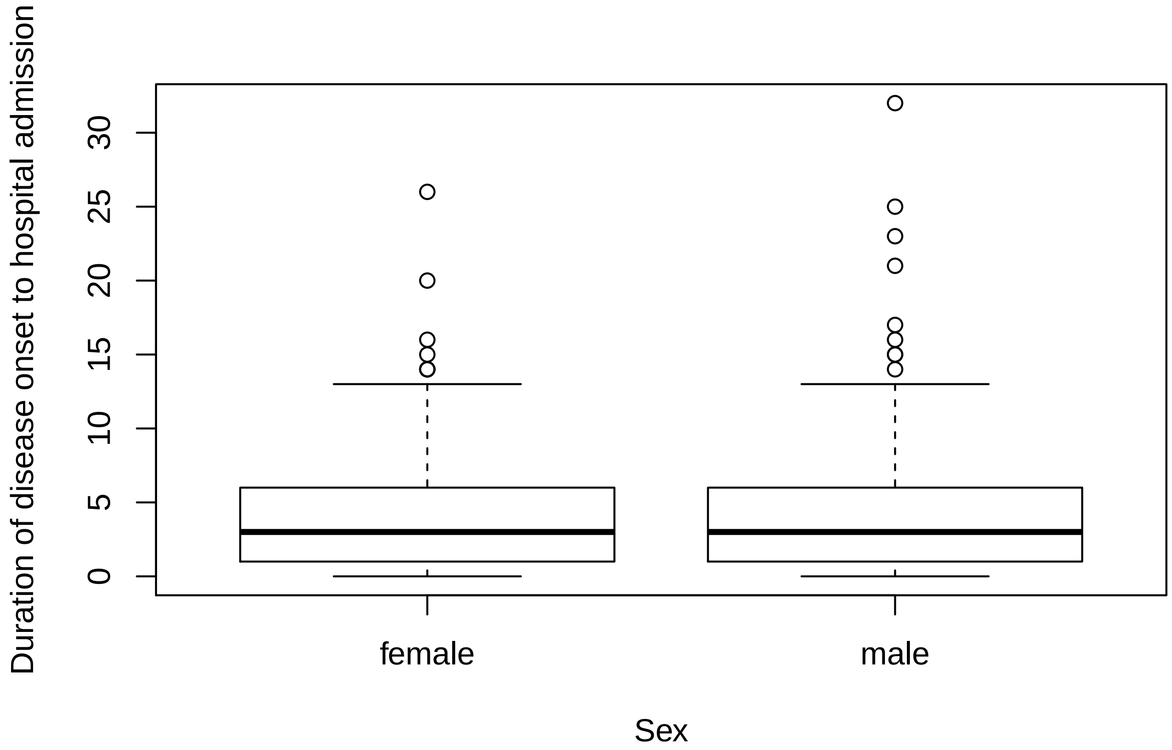
```
##
##          4000 female Female   male   Male    N/A
##      4      0    238      4    323      3      0
```

Use same encoding for male and female

```
indiv_data$sex [indiv_data$sex %in% c("female","Female")] = "female"
indiv_data$sex [indiv_data$sex %in% c("male","Male")] ="male"
indiv_data$sex [!(indiv_data$sex %in% c("male","female"))] =NA
indiv_data$sex = factor(indiv_data$sex)
```

Create a boxplot of sex versus duration of disease onset to hospital

```
boxplot(onset_to_hospital_ad ~sex, data = indiv_data, xlab = "Sex", ylab = "Duration of disease onset t
```



Fit the logistic regression model

```
fit <- glm(sex ~ onset_to_hospital_ad, data = indiv_data, family=binomial)
```

Odd ratios are close to 1 (Estimate 0.01103). The duration of disease onset to hospital admission are comparable between males and females. No evidence that duration is shorter for male or female.

```
summary(fit)
```

```
##
## Call:
## glm(formula = sex ~ onset_to_hospital_ad, family = binomial,
##     data = indiv_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.414  -1.297   1.036   1.062   1.071
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.25452    0.11690   2.177   0.0295 *
## onset_to_hospital_ad  0.01103    0.02051   0.538   0.5905
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 774.95  on 567  degrees of freedom
## Residual deviance: 774.66  on 566  degrees of freedom
##   (13550 observations deleted due to missingness)
## AIC: 778.66
##
## Number of Fisher Scoring iterations: 4
```

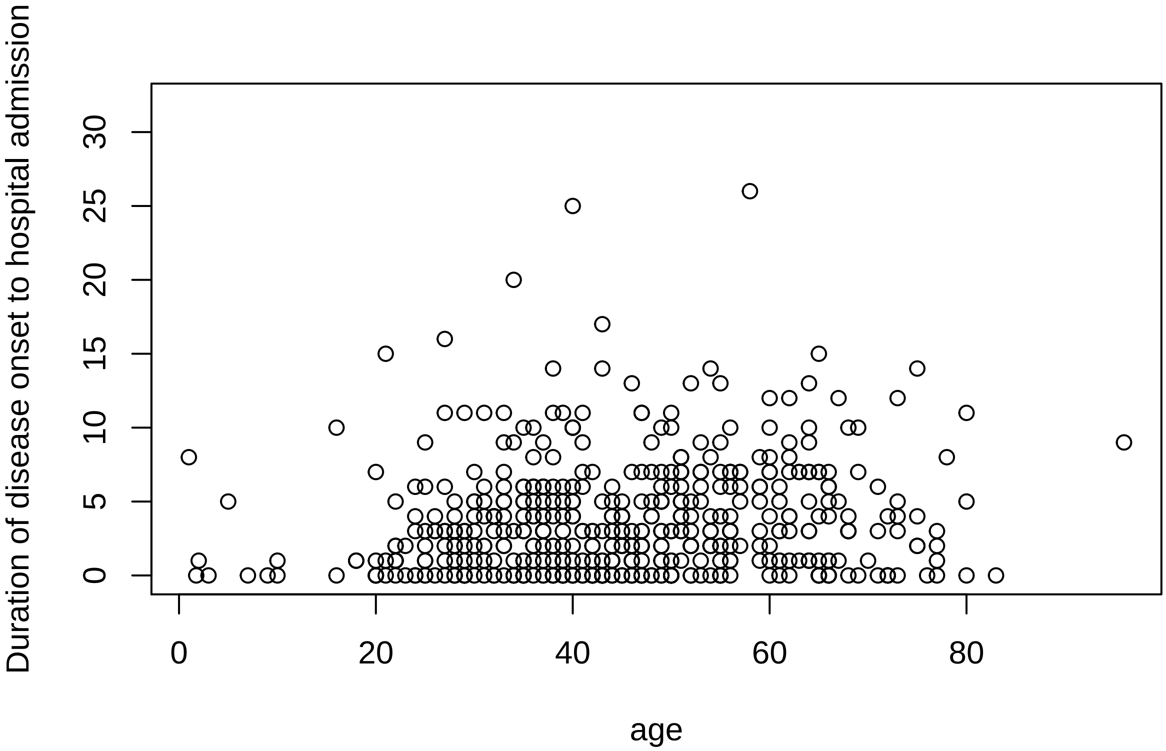
Now we will perform linear regression. Our variable of interest being age We need first to put the variable in the correct type. A warning message will appear for inconsistent entries.

```
indiv_data$age = as.numeric(as.character(indiv_data$age))
```

```
## Warning: NAs introduced by coercion
```

Create a scatterplot

```
plot(onset_to_hospital_ad ~ age, data = indiv_data, xlab = "age", ylab = "Duration of disease onset to 
```

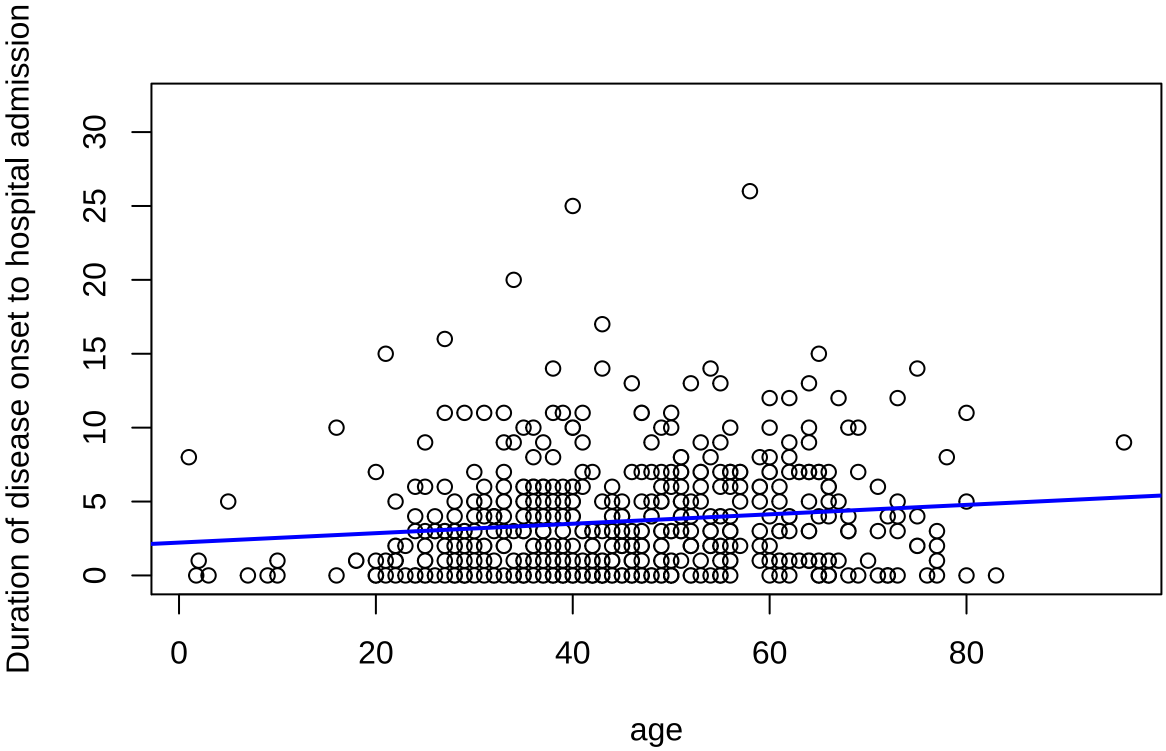


Fit the linear regression model

```
fit <- lm(onset_to_hospital_ad ~ age, data = indiv_data)
```

Add the fitted line to the scatterplot

```
plot(onset_to_hospital_ad ~ age, data = indiv_data, xlab = "age", ylab = "Duration of disease onset to h
abline(fit, col = "blue", lwd=2)
```



There is some evidence that duration of disease onset to hospital admission increases with age

```
summary(fit)
```

```
##
## Call:
## lm(formula = onset_to_hospital_ad ~ age, data = indiv_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8688 -2.9199 -0.9227  1.8941 21.9288
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.22077    0.52714   4.213    3e-05 ***
## age          0.03190    0.01107   2.882  0.00412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.799 on 496 degrees of freedom
##   (13620 observations deleted due to missingness)
## Multiple R-squared:  0.01647,    Adjusted R-squared:  0.01449
## F-statistic: 8.307 on 1 and 496 DF,  p-value: 0.004121
```