PASSI ANALYSIS

Step by Passi:

- Angeletti Giacomo
- Crapanzano Edoardo
- Dumitru Vlad Adrian
- Zago Francesco

Introduction and curiosities

PASSI = Progressi delle Aziende Sanitarie per la Salute in Italia
(Progress of Health Authorities for Health in Italy)

It is a national health surveillance system in Italy that monitors the health behaviors and risk factors of the adults aged 18-69.

A large survey and monitoring program that regularly collects information from people across Italy about things like smoking, physical activity, alcohol use, weight.

This information hepls public health authorities make better decisions and create effective programs to improve the health of the population.

The PASSI file used for the project contain approximately 500.000 observations, from the year 2008 to 2020.

We decided to analyze in particular the years 2009 and 2019 to see how the situation has changed during the decade.

Table of contents

01

Socioeconomic Variables related to Chronic Diseases

02

Cardiovascular Risk in relation to the quantity of Cigarettes smoked and Alcohol consumption

03

Hypertension Conditions related to Physical Activity

04

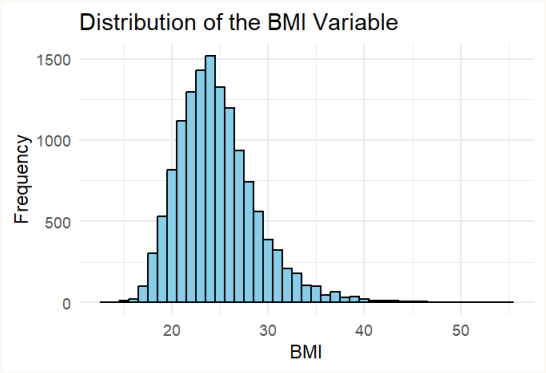
Physical Health in relation to Mental Well-Being

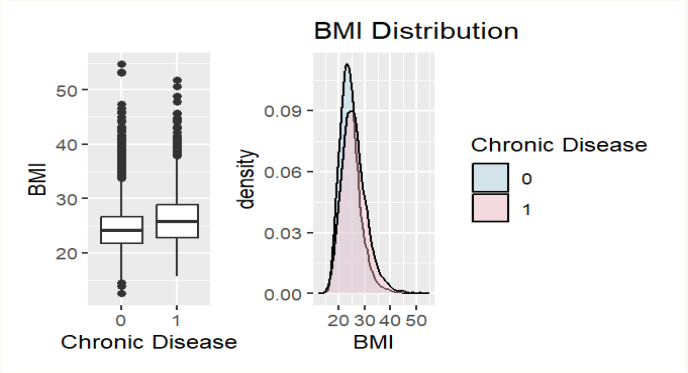
01

Socioeconomic Variables related to Chronic Diseases

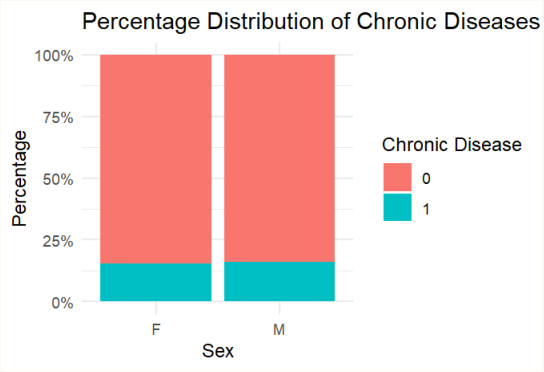
Core Variables

BMI

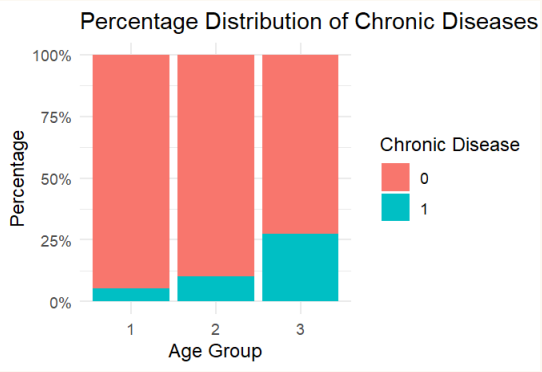




Sex

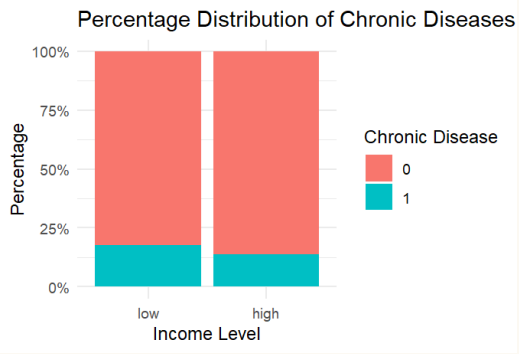


Age Group



Additional Variables

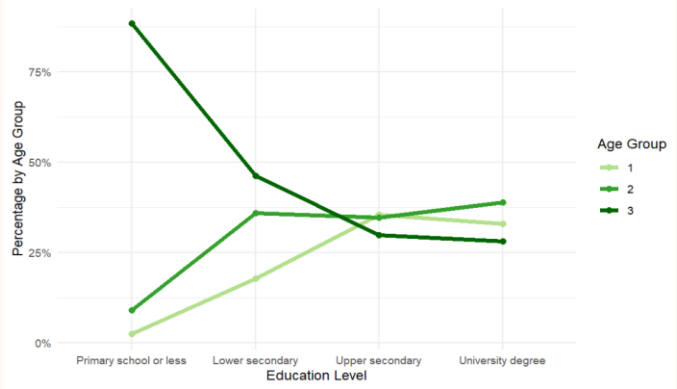
Income Level



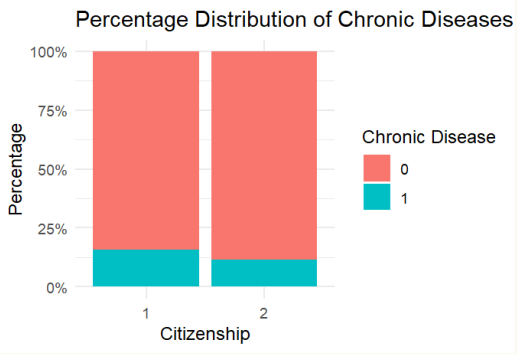

```
Pearson's Chi-squared test with Yates' continuity correction

data:  table(dtpa$s14_disp_econ_mod, dtpa$malattiecroniche)
X-squared = 35.071, df = 1, p-value = 3.179e-09
```

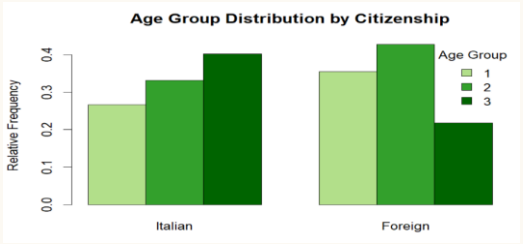
Education Level

	0	1
Primary school or less	0.6126205	0.3873795
Lower secondary	0.8161601	0.1838399
Upper secondary	0.8872059	0.1127941
University degree	0.8915254	0.1084746



Citizenship





Logistic Regression Analysis

$$\text{malattiecroniche}_i = \begin{cases} 1 & \text{if the subject has at least one chronic disease} \\ 0 & \text{otherwise} \end{cases}$$

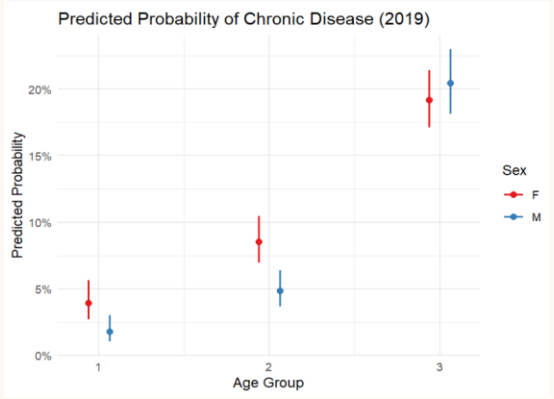
Best Model	2009	2019
BMI	0.04 (4.3%) ***	0.05 (4.9%) ***
Sex	0.08	-0.80 (-55%) *
Age Group 35-49	0.48 (61%) ***	0.80 (122%) ***
Age Group 50-69	1.52 (355%) ***	1.64 (414%) ***
Education Level (LowS, UpS, Uni)	-0.43 -0.67 -0.56 *** (-35% -49% -43%)	-0.54 -0.75 -0.75 *** (-42% -53% -53%)

2009

```
s14_disp_econ_modhigh      **
```

2019

```
sesso_intervistatoM              -0.80230    0.33215  -2.415 0.015714 *
sesso_intervistatoM:claeta33      0.91681    0.34872   2.629 0.008561 **
```

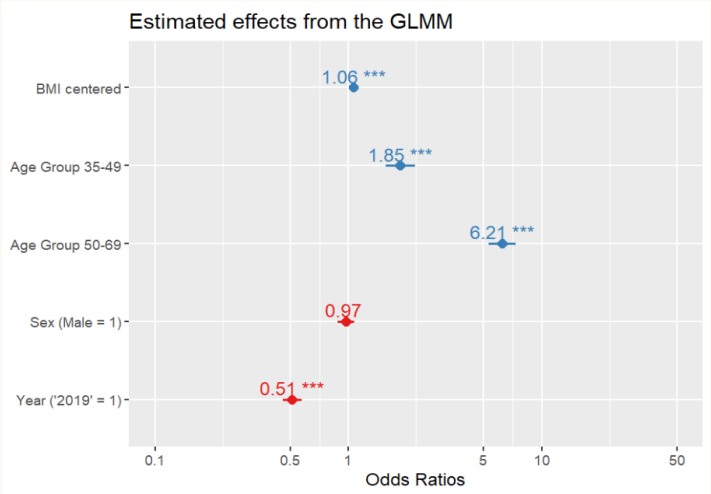


GLMM

Year as fixed effect

ASL as random effect

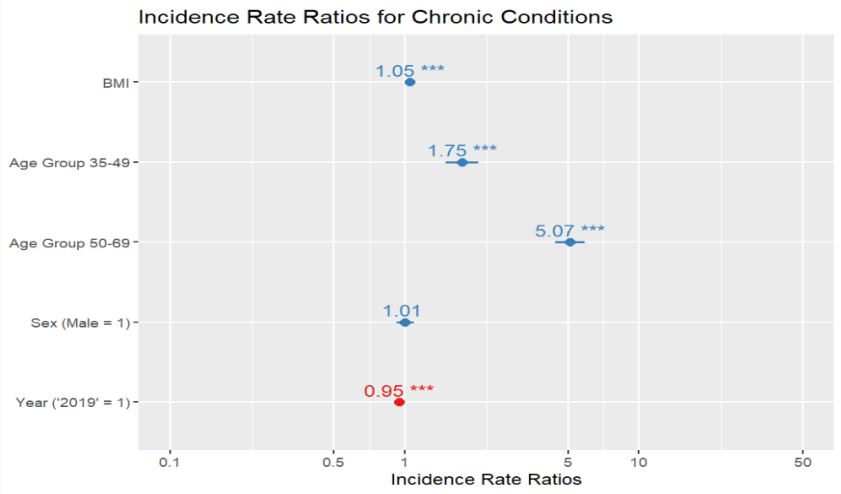
```
# Random Effects

Within-Group Variance          3.29 (1.81)
Between-Group Variance
  Random Intercept (asl)       0.03 (0.16)
N (groups per factor)
  asl                           209
Observations                  13459
```



Count Data Model

$$Y_i = D_i + R_i + B_i + M_i + T_i$$

$$X_i = \begin{cases} 1 & \text{if individual } i \text{ has the condition} \\ 0 & \text{otherwise} \end{cases}$$



02

Cardiovascular Risk in relation to the quantity of Cigarettes and Alcohol Consumption

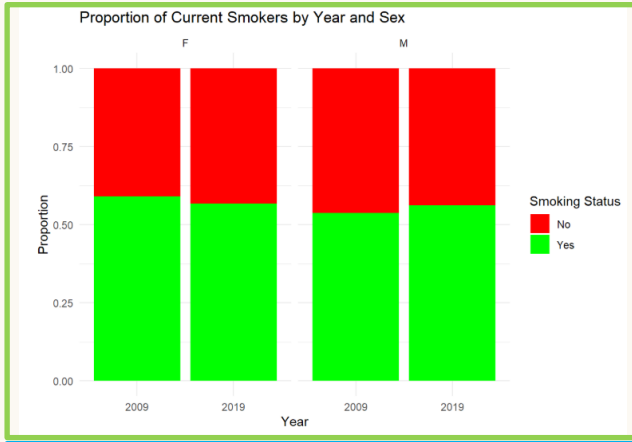
Core variables

This part of the project focuses on how risk factors, specifically cigarettes and alcohol consumption, can impact cardiovascular risk and, if so, how significantly they impact.

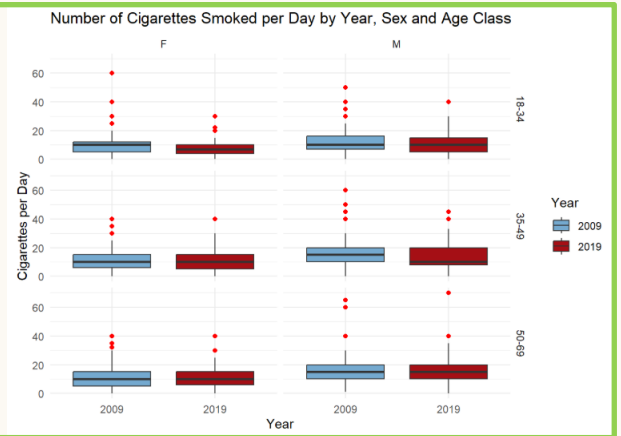
Points of analysis

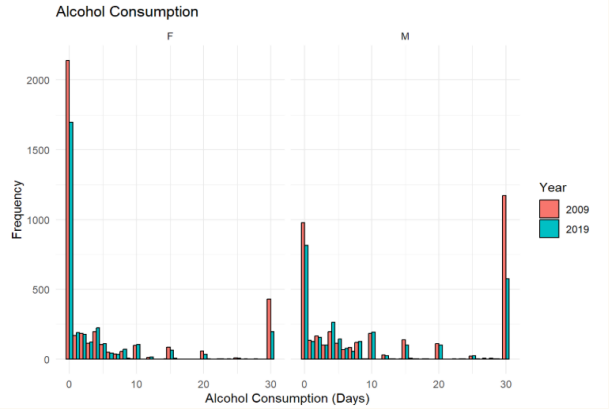
- **Cigarette Consumption:** analysing the smoking status and then, for the smokers, the number of cigarettes smoked per day;
- **Alcohol Consumption:** how many days did the subject drink at least one unit of alcoholic beverage during the last 30 days and, on drinking days, the average quantities of units of alcoholic beverages in a day;
- **Cardiovascular Risk:** how cardiovascular risk varies by sex, year, age class and territorial level (ASL and region);
- **Impact of Cigarettes and Alcohol:** with a Generalized Linear Mixed Model (GLMM) to understand how the factors analysed in the points above are correlated with each other. In particular, analyse whether cigarettes and alcohol can have a significant impact on cardiovascular risk.

Core variables

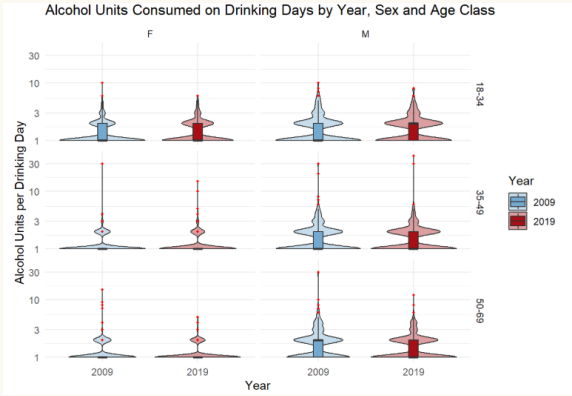


Cigarettes





Alcohol



Impact of Cigarettes and Alcohol

```r
# Model with ASL as random intercept
Model_Card_ASL = glmer(cardiovasc ~
                  as.factor(s03_fumo_att) * as.factor(anno) +
                  s05_alcool_gg * as.factor(anno)+
                  as.factor(sesso_intervistato) +
                  as.factor(claeta3) +
                  (1 | asl),
                data = Cigarettes_Alcohol,
                family = binomial,
                control = glmerControl(optimizer = "bobyqa"))

# Model with Region as random intercept
Model_Card_Regione = glmer(cardiovasc ~
                  as.factor(s03_fumo_att) * as.factor(anno) +
                  s05_alcool_gg * as.factor(anno)+
                  as.factor(sesso_intervistato) +
                  as.factor(claeta3) +
                  (1 | regione),
                data = Cigarettes_Alcohol,
                family = binomial,
                control = glmerControl(optimizer = "bobyqa"))

# Let's compare the two models
anova(Model_Card_ASL, Model_Card_Regione)
```

```r
parameters::random_parameters(Model_Card_ASL)
parameters::random_parameters(Model_Card_Regione)
```

Since the model includes a random intercept for territorial level (ASL or region) we need to see which of the two variables is more significant

```
## Data: Cigarettes_Alcohol
## Models:
## Model_Card_ASL: cardiovasc ~ as.factor(s03_fumo_att) * as.factor(anno) + s05_alcool_gg * as.factor(ann
o) + as.factor(sesso_intervistato) + as.factor(claeta3) + (1 | asl)
## Model_Card_Regione: cardiovasc ~ as.factor(s03_fumo_att) * as.factor(anno) + s05_alcool_gg * as.factor
(anno) + as.factor(sesso_intervistato) + as.factor(claeta3) + (1 | regione)
##                  npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## Model_Card_ASL     10 2425.9 2493.2 -1202.9    2405.9
## Model_Card_Regione 10 2423.7 2491.0 -1201.8    2403.7 2.1892  0
```

```
## # Random Effects
##
## Within-Group Variance     3.29 (1.81)
## Between-Group Variance
##   Random Intercept (asl)  0.08 (0.28)
## N (groups per factor)
##   asl                      209
## Observations              6178
```

```
## # Random Effects
##
## Within-Group Variance         3.29 (1.81)
## Between-Group Variance
##   Random Intercept (regione)  0.04 (0.21)
## N (groups per factor)
##   regione                      21
## Observations                 6178
```

Values are quite similar but we prefer the Model with Region as random intercept

Impact of Cigarettes and Alcohol

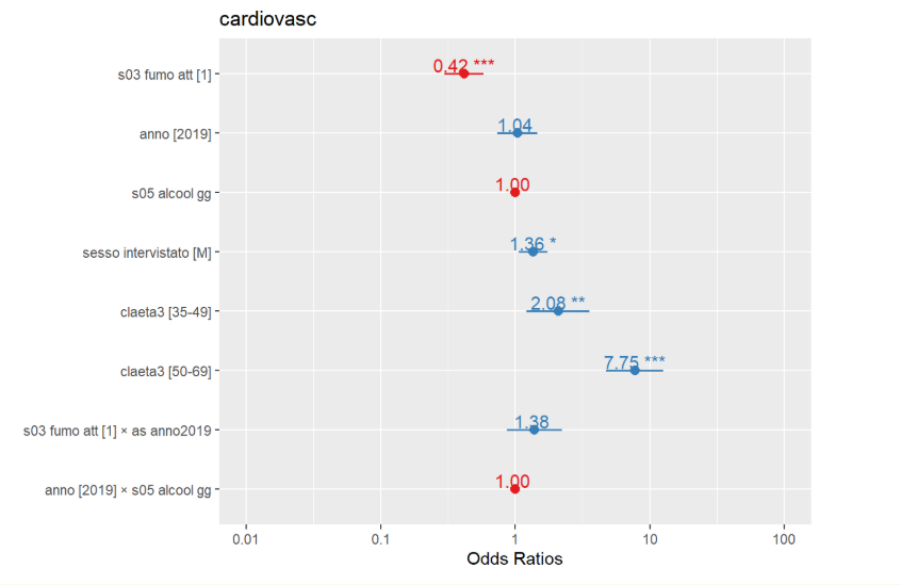
①

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## cardiovasc ~ as.factor(s03_fumo_att) * as.factor(anno) + s05_alcool_gg *
##     as.factor(anno) + as.factor(sesso_intervistato) + as.factor(claeta3) +
##     (1 | regione)
##    Data: Cigarettes_Alcohol
## Control: glmerControl(optimizer = "bobyqa")
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    2423.7    2491.0   -1201.8    2403.7      6168
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.4777 -0.2926 -0.1758 -0.1170 12.8731
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  regione (Intercept) 0.0435   0.2086
## Number of obs: 6178, groups:  regione, 21
##
```

②

```
## Fixed effects:
##                                             Estimate Std. Error z value
## (Intercept)                                -4.064944   0.276784 -14.686
## as.factor(s03_fumo_att)1                   -0.876969   0.171099  -5.125
## as.factor(anno)2019                         0.035602   0.174057   0.205
## s05_alcool_gg                              -0.002641   0.005924  -0.446
## as.factor(sesso_intervistato)M              0.306223   0.125360   2.443
## as.factor(claeta3)35-49                     0.733750   0.273961   2.678
## as.factor(claeta3)50-69                     2.047800   0.249195   8.218
## as.factor(s03_fumo_att)1:as.factor(anno)2019 0.324744   0.241177   1.346
## as.factor(anno)2019:s05_alcool_gg          -0.004570   0.009085  -0.503
##                                            Pr(>|z|)
## (Intercept)                                 < 2e-16 ***
## as.factor(s03_fumo_att)1                   2.97e-07 ***
## as.factor(anno)2019                         0.8379
## s05_alcool_gg                               0.6558
## as.factor(sesso_intervistato)M              0.0146 *
## as.factor(claeta3)35-49                     0.0074 **
## as.factor(claeta3)50-69                     < 2e-16 ***
## as.factor(s03_fumo_att)1:as.factor(anno)2019  0.1781
## as.factor(anno)2019:s05_alcool_gg           0.6149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) as.(03__)1 as.()2019 s05_l_ a.(_)M a.(3)3 a.(3)5 a.(03__)1:
## as.f(03__)1 -0.297
## as.fc()2019 -0.283  0.299
## s05_alcl_gg -0.155  0.013      0.420
## as.fctr(_)M -0.264  0.069      0.017    -0.225
## as.(3)35-49 -0.753  0.055     -0.014    -0.030  0.006
## as.(3)50-69 -0.834  0.112     -0.025    -0.065  0.001  0.836
## a.(03__)1:.  0.152 -0.696     -0.492    -0.021 -0.033 -0.005 -0.015
## a.()2019:05  0.167 -0.026     -0.579    -0.614 -0.002  0.006  0.009  0.036
```

Impact of Cigarettes and Alcohol



Looking at the results of the model:

- ***Age classes***: 675% (for 50-69 group) and 108% (for 35-49 group) higher probability of having cardiovascula risk;
- ***Interactions***: both for cigarettes and alcohol are not significant;
- ***Alcohol***: no correlation between alcohol and cardiovascular risk -> it is not statistically significant;
- ***Smoking***: very strange situation because it is statistically significant but there is an inverse correlation -> smokers are more likely to have a low cardiovascular risk than non-smokers.

New variable: Smoking History

The subject, in his entire life, has smoked at least 100 cigarettes in total?

```r
# Generalized Linear Mixed Model with smoking history
Model_Card_Regione_2 = glmer(cardiovasc ~
                  as.factor(s03_fumo) * as.factor(anno) +
                  s05_alcool_gg * as.factor(anno)+
                  as.factor(sesso_intervistato) +
                  as.factor(claeta3) +
                  (1 | regione),
                data = Cigarettes_Alcohol,
                family = binomial,
                control = glmerControl(optimizer = "bobyqa"))
summary(Model_Card_Regione_2)
```

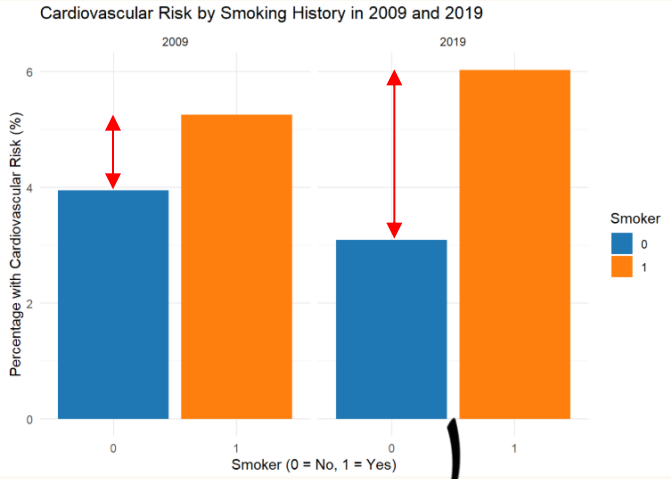
①

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cardiovasc ~ as.factor(s03_fumo) * as.factor(anno) + s05_alcool_gg *
##     as.factor(anno) + as.factor(sesso_intervistato) + as.factor(claeta3) +
##     (1 | regione)
##    Data: Cigarettes_Alcohol
## Control: glmerControl(optimizer = "bobyqa")
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    4469.1    4544.1   -2224.6    4449.1     13300
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.4300 -0.2819 -0.1472 -0.1091 13.3617
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  regione (Intercept) 0.02825  0.1681
## Number of obs: 13310, groups:  regione, 21
##
```

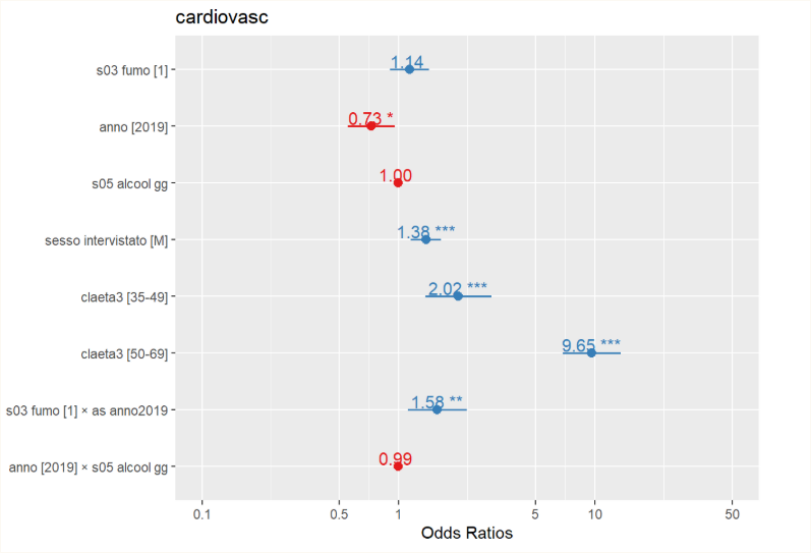
②

```
## Fixed effects:
##                                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                -4.720553   0.189207 -24.949  < 2e-16
## as.factor(s03_fumo)1                        0.128320   0.117153   1.095 0.273378
## as.factor(anno)2019                        -0.317521   0.140117  -2.266 0.023445
## s05_alcool_gg                              -0.002826   0.004564  -0.619 0.535825
## as.factor(sesso_intervistato)M              0.323135   0.091256   3.541 0.000399
## as.factor(claeta3)35-49                     0.703422   0.196111   3.587 0.000335
## as.factor(claeta3)50-69                     2.266789   0.173045  13.099  < 2e-16
## as.factor(s03_fumo)1:as.factor(anno)2019    0.454884   0.175601   2.590 0.009585
## as.factor(anno)2019:s05_alcool_gg          -0.005931   0.007323  -0.810 0.418016
##
## (Intercept)                                ***
## as.factor(s03_fumo)1
## as.factor(anno)2019                        *
## s05_alcool_gg
## as.factor(sesso_intervistato)M             ***
## as.factor(claeta3)35-49                    ***
## as.factor(claeta3)50-69                    ***
## as.factor(s03_fumo)1:as.factor(anno)2019 **
## as.factor(anno)2019:s05_alcool_gg
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) as.(03_)1 as.()2019 s05_l_ a.(_)M a.(3)3 a.(3)5 a.(03_)1:
## as.fc(03_)1 -0.249
## as.fc()2019 -0.301  0.394
## s05_alcl_gg -0.092 -0.135     0.259
## as.fctr(_)M -0.197 -0.148    -0.028    -0.219
## as.(3)35-49 -0.732 -0.004     0.000    -0.036  0.019
## as.(3)50-69 -0.820 -0.022    -0.028    -0.078  0.045  0.809
## a.(03_)1:.(  0.192 -0.655    -0.638     0.107  0.022 -0.011 -0.001
## a.()2019:05  0.111  0.108    -0.341    -0.584 -0.016  0.009  0.011 -0.168
```

New variable: Smoking History



In 2019 the difference is more marked

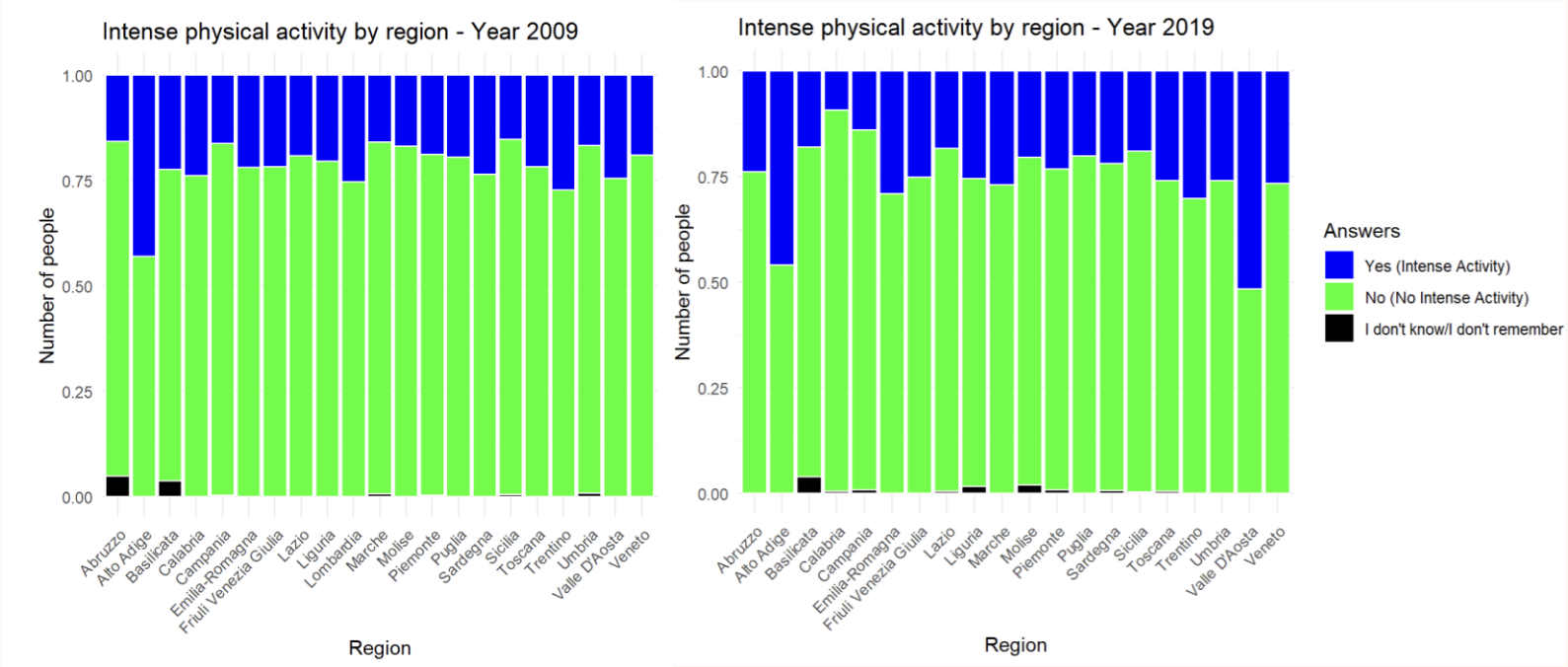


- **The interaction between smoking status and year 2019 is statistically significant**.
- 58% higher cardiovascular risk that non-smokers.
- In 2019 being a smoker significantly increased the probability of having a cardiovascular condition, more than in 2009.

03

Hypertension Conditions related to Physical Activity

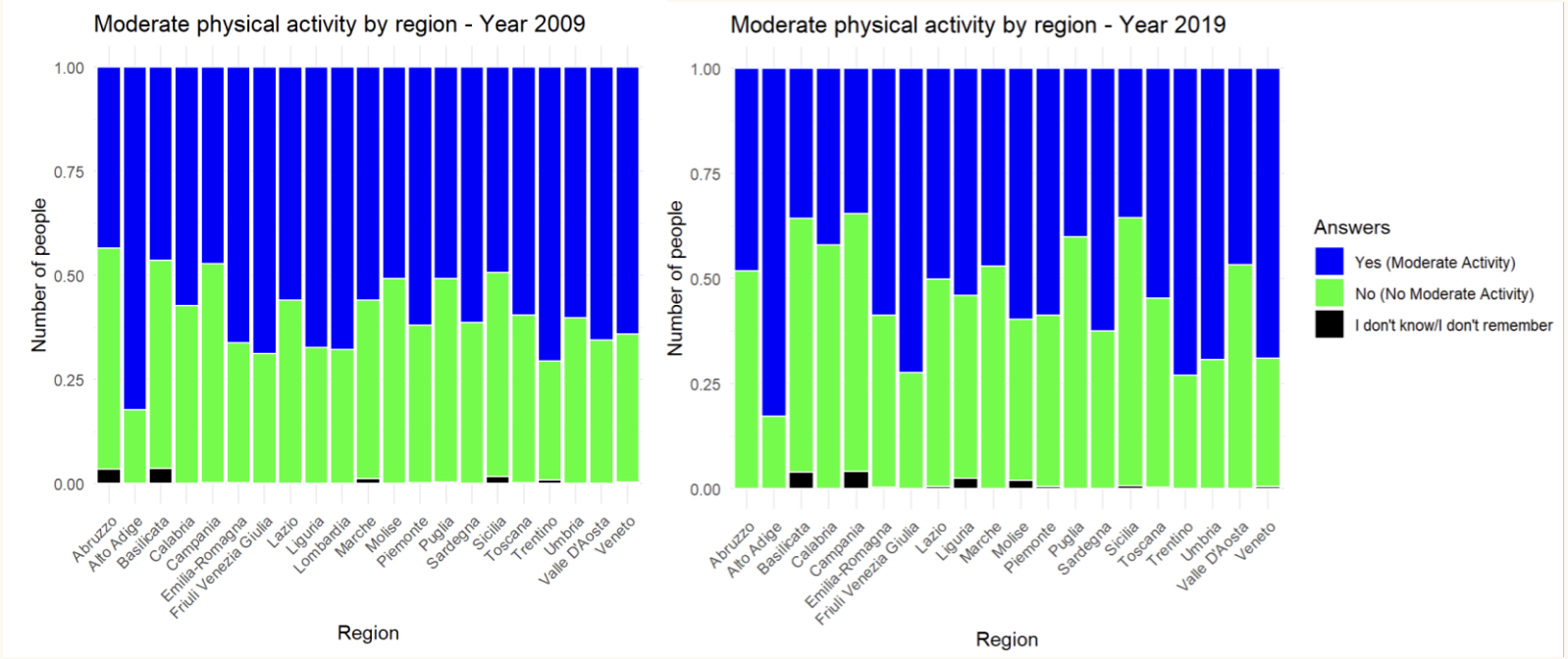
Intense Physical Activity Distribution




```
##   attivita_label                    n percentuale
##   <fct>                         <int>       <dbl>
## 1 Yes (Intense Activity)         1499        20.1
## 2 No (No Intense Activity)       5938        79.6
## 3 I don't know/I don't remember    22        0.29
```

```
##   attivita_label                    n percentuale
##   <fct>                         <int>       <dbl>
## 1 Yes (Intense Activity)         1430        23.4
## 2 No (No Intense Activity)       4660        76.2
## 3 I don't know/I don't remember    28        0.46
```

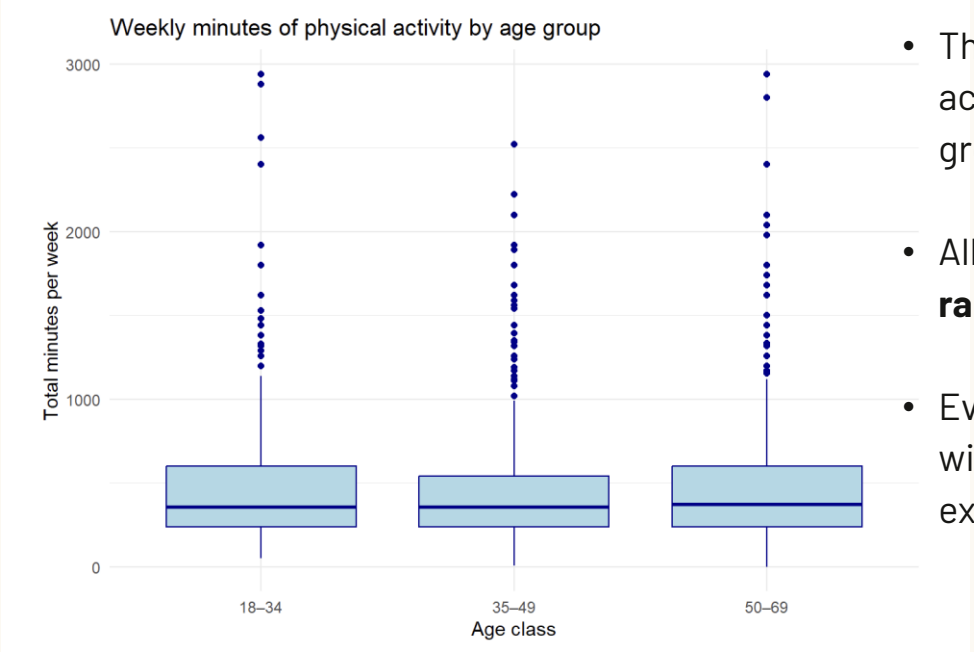
Moderate Physical Activity Distribution




```
##   attivita_label                      n percentuale
##   <fct>                           <int>       <dbl>
## 1 Yes (Moderate Activity)          4468        59.9
## 2 No (No Moderate Activity)        2971        39.8
## 3 I don't know/I don't remember      25        0.33
```

```
##   attivita_label                      n percentuale
##   <fct>                           <int>       <dbl>
## 1 Yes (Moderate Activity)          3302        54.0
## 2 No (No Moderate Activity)        2774        45.3
## 3 I don't know/I don't remember      43         0.7
```

Boxplot of Weekly Minutes of Physical Activity



- The median weekly minutes of physical activity is fairly similar across all age groups (18-34, 35-49, 50-69);
- All groups show a wide **interquartile range (IQR)** and long whiskers;
- Every age group has several outliers, with some individuals reporting extremely high levels of weekly activity.

Hypertension and Medical Advice

```r
# Filter the responses without including NA values
dataset_contingency = dataset_definitivo %>%
  filter(s07_press_alta %in% c(1, 2),
         s07_sugg_att_fis_press %in% c(1, 2)) %>%
  mutate(
    ipertensione = ifelse(s07_press_alta == 1, "Hypertensive", "No Hypertensive"),
    consiglio_att_fisica = ifelse(s07_sugg_att_fis_press == 1, "Advice Received", "No Advice")
  )

# Absolute frequency table
contingency_table = table(dataset_contingency$ipertensione, dataset_contingency$consiglio_att_fis
print(contingency_table)
```

```
##
##                Advice Received No Advice
##   Hypertensive            1989       436
```

Hide

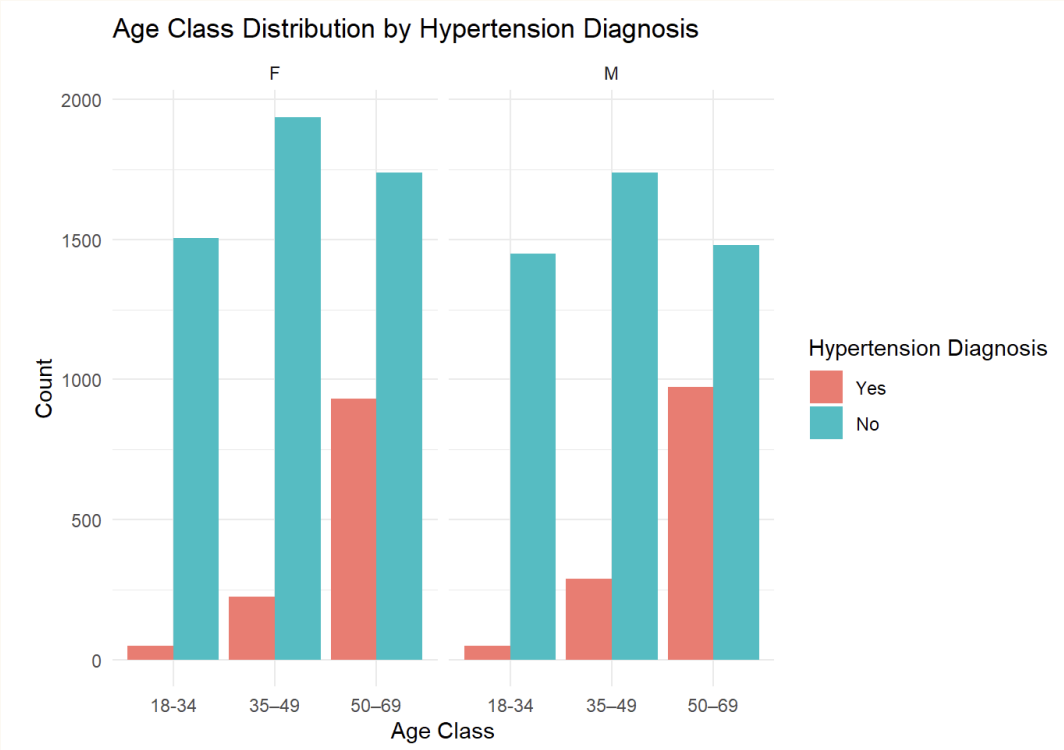
```r
# Percentages per raw
percentage_table <- prop.table(contingency_table, margin = 1) * 100
print(round(percentage_table, 1))
```

```
##
##                Advice Received No Advice
##   Hypertensive              82         18
```

Contingency table

Among those who replied to the survey, 82% of people who was diagnosed with hypertension was also suggested to practice more physical activity.

Hypertension Distribution by Age and Gender



- Hypertension becomes significantly more common with age;
- Women aged 35-49 appear to be less affected than men in the same group.

Generalized Linear Model on Hypertension

```
##
## Call:
## glm(formula = s07_press_alta_bin ~ claeta3 + sesso_intervistato,
##     family = binomial(link = "logit"), data = dataset_filtered_press)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.98099    0.11428 -43.587  < 2e-16 ***
## claeta3              1.44655    0.04110  35.194  < 2e-16 ***
## sesso_intervistato1  0.23792    0.04801   4.955 7.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12494  on 12366  degrees of freedom
## Residual deviance: 10722  on 12364  degrees of freedom
## AIC: 10728
##
## Number of Fisher Scoring iterations: 5
```

- The **log-odds** of having high blood pressure for an individual in the **18-34 age group** and for the individual being **female** is **-4.98**, which corresponds to a very **low probability** of hypertension;

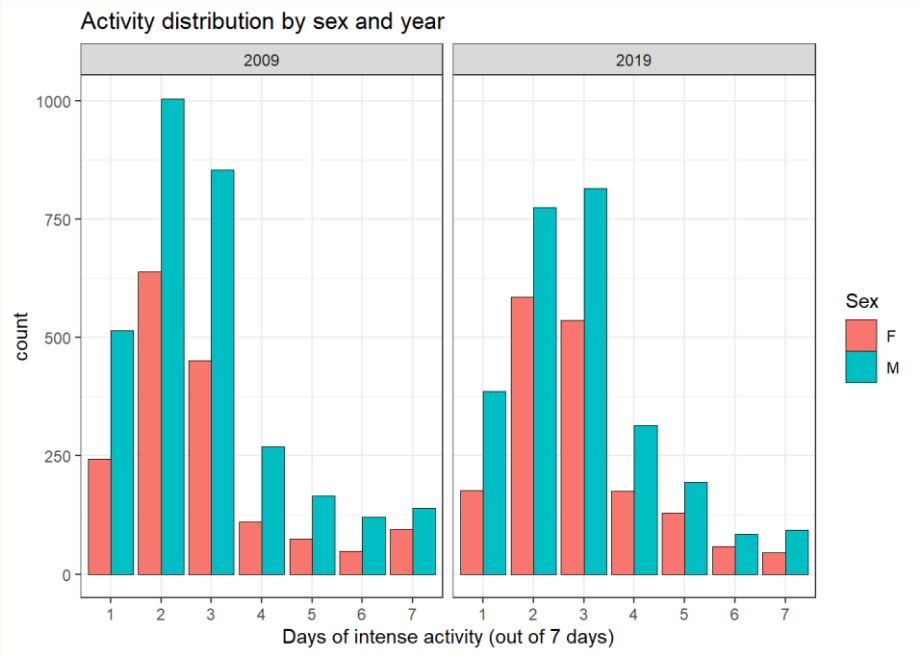
- Each older age group is associated with **over four times greater odds** of having blood pressure.

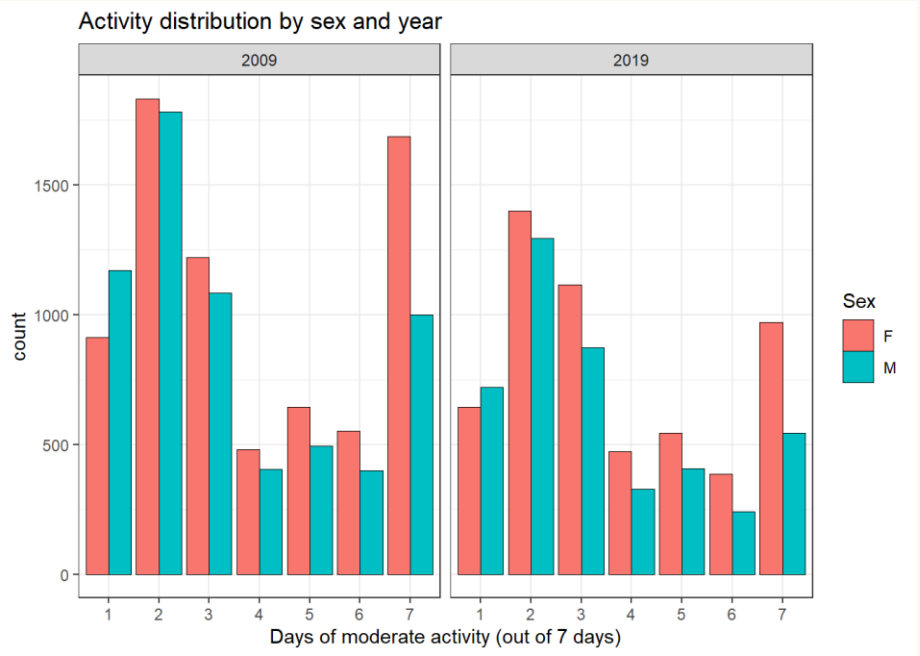
04

Physical Health in relation to Mental Well-Being

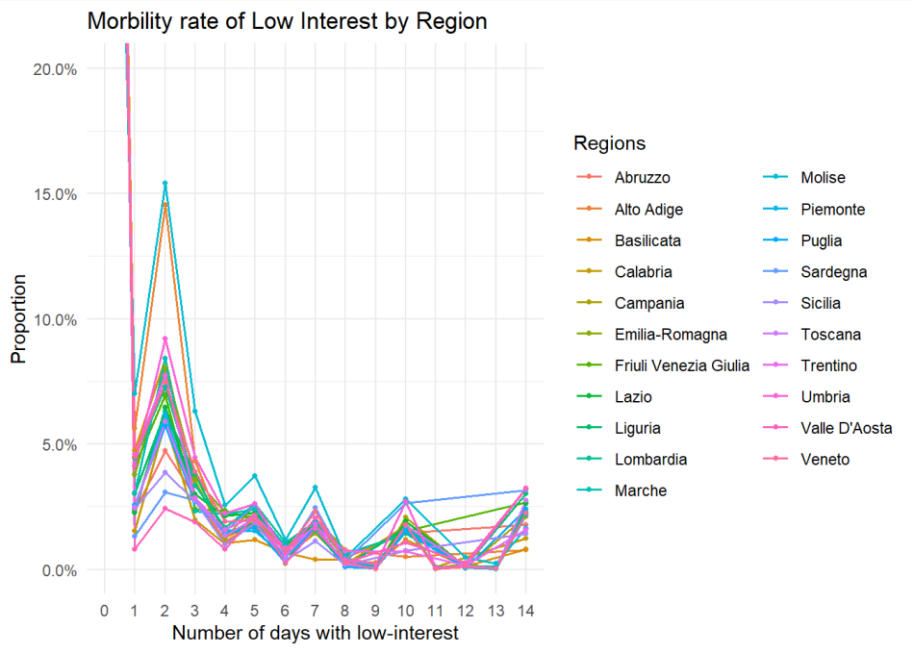
Explorative Analysis

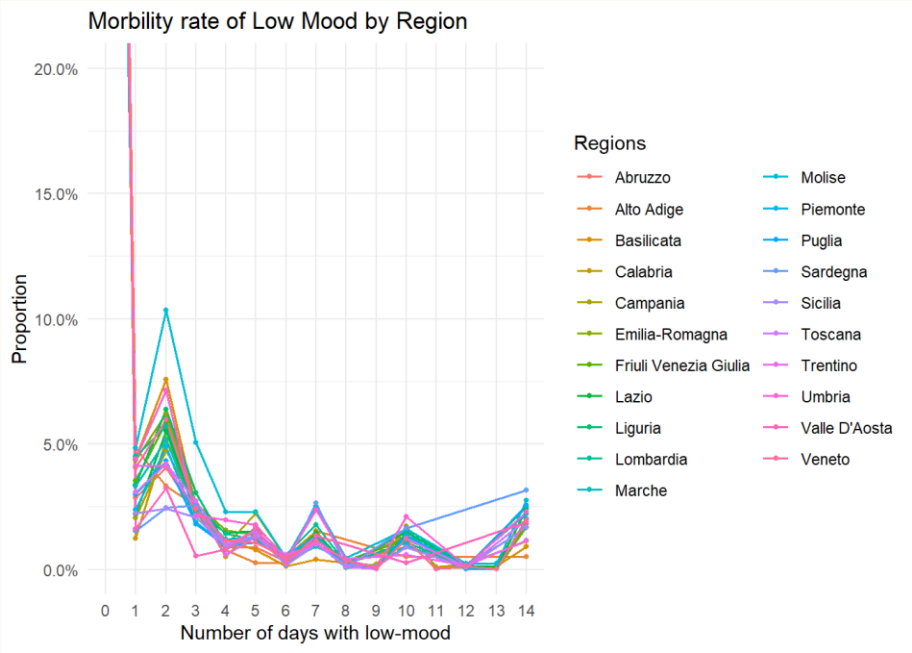
Plots of the distribution of intense and moderate activity in days



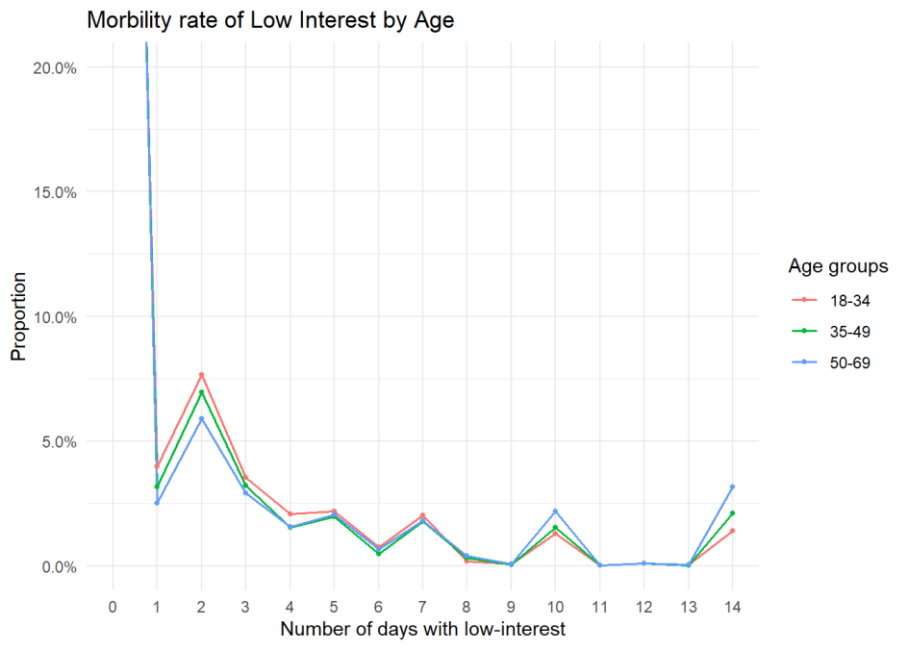


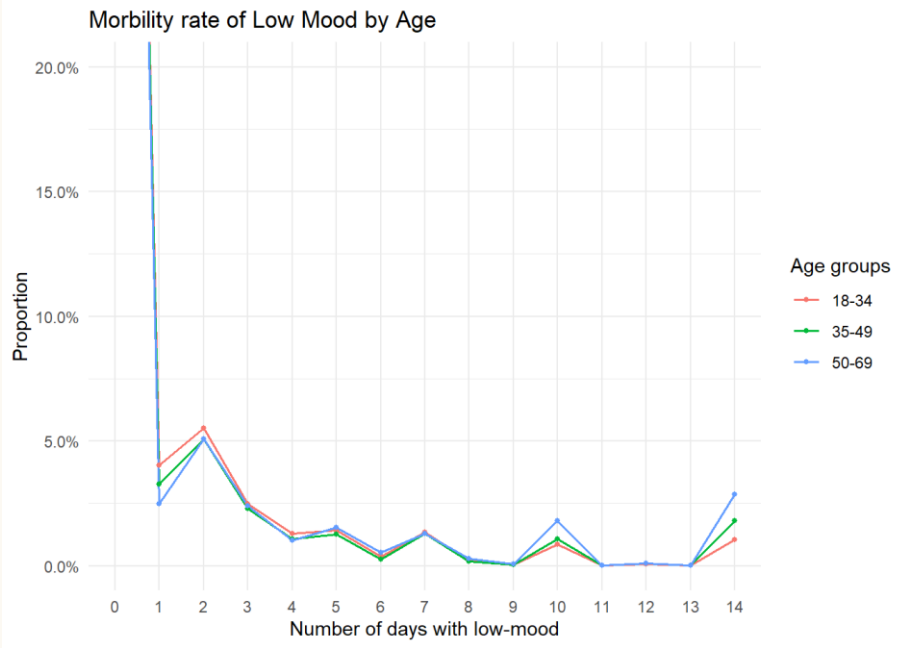
Plots of the proportions/morbidity rates of `s12_poco_inter_gg` and `s12_giu_morale_gg`
- grouping by region and zooming in the interval between 0% and 20% -



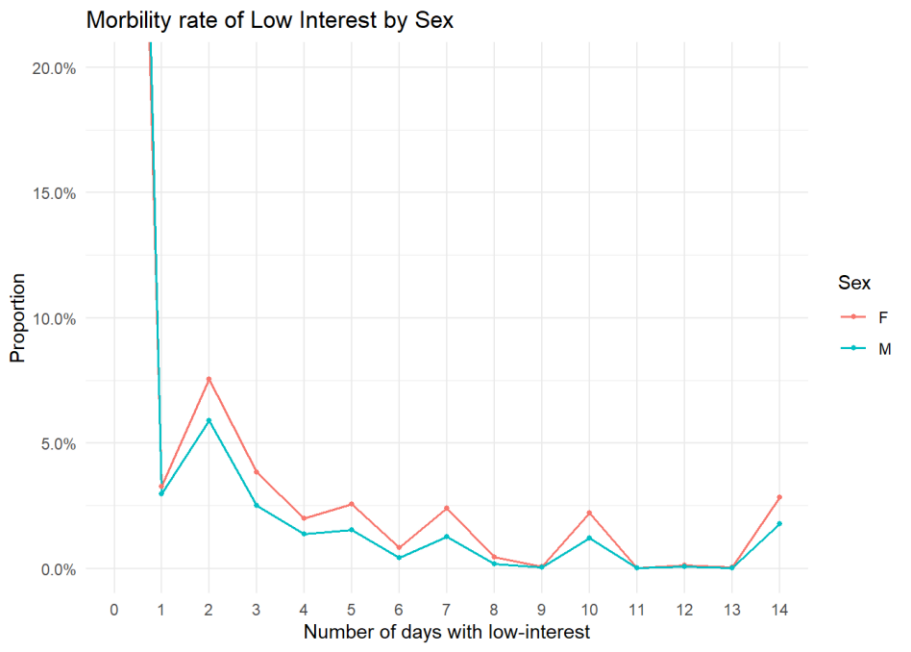


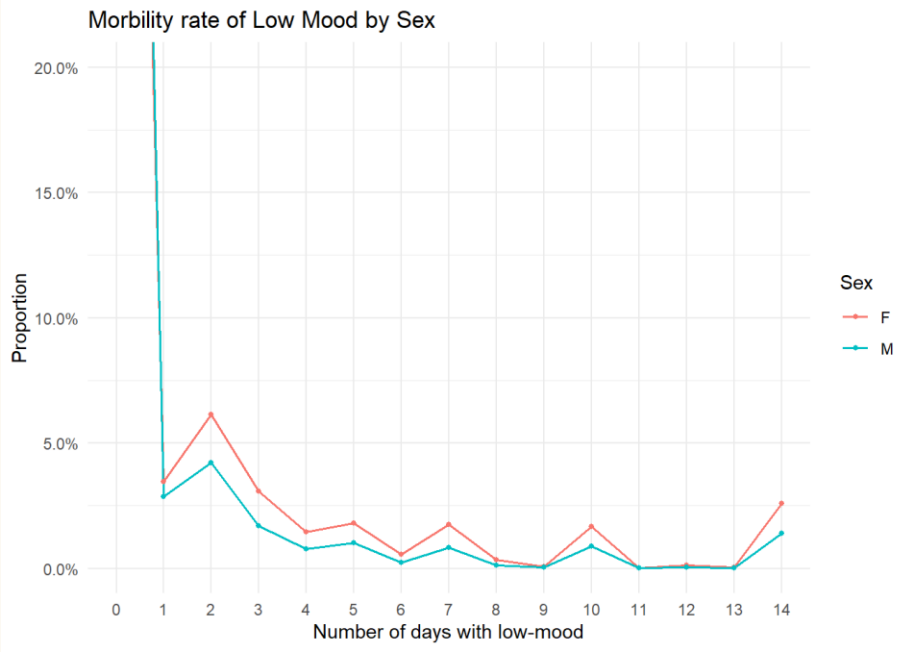
Plots of the proportions/morbidity rates of `s12_poco_inter_gg` and `s12_giu_morale_gg`
- grouping by age and zooming in the interval between 0% and 20% -





Plots of the proportions/morbidity rates of `s12_poco_inter_gg` and `s12_giu_morale_gg`
- grouping by sex and zooming in the interval between 0% and 20% -





Fitting regression models

With `s12_poco_inter_gg` as dependent variable

Baseline: a female 18–34 year-old with 0 days of moderate or intense activity and without daily fruit consumption.

There is a modest positive correlation (0.24) between `s02_tlib_attiv_inte_gg` and `s02_attiv_moder_gg`, but it isn't very strong. With p-value < 2.2e-16, the null hypothesis (correlation = 0) is rejected with strong evidence, so the correlation is statistically significant and we are 95% confident that the true correlation falls in the range between 0.21 and 0.26.

The interaction of the numerical predictors significantly improves the model's fit and should be considered: it suggests that the association between `s02_attiv_moder_gg` and `s12_poco_inter_gg` varies depending on `s02_tlib_attiv_inte_gg`.

```
## Coefficients:
##                                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                  1.256230   0.090309  13.910  < 2e-16
## s02_tlib_attiv_inte_gg                      -0.091365   0.020332  -4.494 7.00e-06
## s02_attiv_moder_gg                          -0.023700   0.013700  -1.730  0.08364
## s04_consu_frutta1-2                         -0.570315   0.071641  -7.961 1.71e-15
## s04_consu_frutta3-4                         -0.780457   0.072764 -10.726  < 2e-16
## s04_consu_frutta5+                          -0.655934   0.080379  -8.160 3.34e-16
## claeta335-49                                -0.188928   0.030335  -6.228 4.72e-10
## claeta350-69                                -0.388337   0.037701 -10.300  < 2e-16
## sesso_intervistatoM                         -0.387415   0.027390 -14.144  < 2e-16
## s02_tlib_attiv_inte_gg:s02_attiv_moder_gg  0.013121   0.004314   3.041  0.00236
##
## (Intercept)                                 ***
## s02_tlib_attiv_inte_gg                      ***
## s02_attiv_moder_gg                          .
## s04_consu_frutta1-2                         ***
## s04_consu_frutta3-4                         ***
## s04_consu_frutta5+                          ***
## claeta335-49                                ***
## claeta350-69                                ***
## sesso_intervistatoM                         ***
## s02_tlib_attiv_inte_gg:s02_attiv_moder_gg **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 17492  on 5240  degrees of freedom
## Residual deviance: 17007  on 5231  degrees of freedom
## AIC: 21290
##
## Number of Fisher Scoring iterations: 6
```

Fitting regression models

With `s12_giu_morale_gg` as dependent variable

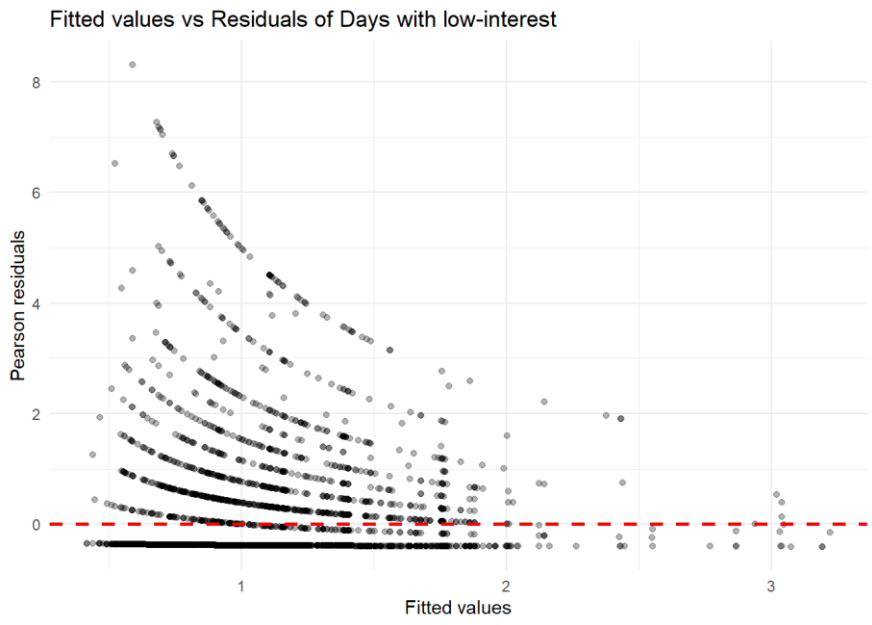
The partial effect of `s02_tlib_attiv_inte_gg` is not statistically significant because p-value = 0.12. Also the partial effect of `s02_attiv_moder_gg` is not statistically significant because p-value = 0.12. Thus, I will take the simpler model with categorical predictors with significant partial effects on the dependent variable.

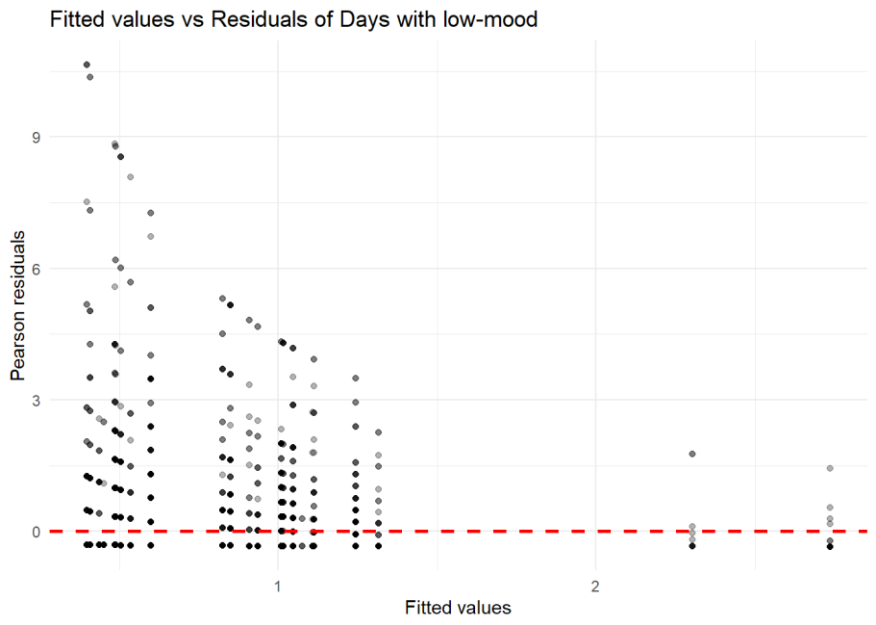
Poisson distribution has identical mean and variance, but in this case both the models have over-dispersion because alpha>1 (dispersion parameter) with strong statistical evidence (p-value < 2.2e-16).

After fitting Negative Binomial Model there is a drastic reduction in over-dispersion that makes this model more appropriate for inference and confidence intervals.

```
## Call:
## glm(formula = s12_giu_morale_gg ~ s04_consu_frutta + claeta3 +
##     sesso_intervistato, family = poisson(link = "log"), data = p
##     filter(if_all(everything(), ~!is.na(.))))
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.94258    0.08175  11.530  < 2e-16 ***
## s04_consu_frutta1-2 -0.77963    0.08278  -9.418  < 2e-16 ***
## s04_consu_frutta3-4 -0.97815    0.08396 -11.650  < 2e-16 ***
## s04_consu_frutta5+  -0.83122    0.09253  -8.983  < 2e-16 ***
## claeta335-49        -0.12740    0.03726  -3.419 0.000628 ***
## claeta350-69        -0.13622    0.04339  -3.140 0.001691 **
## sesso_intervistatoM -0.70597    0.03343 -21.116  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 14388  on 5240  degrees of freedom
## Residual deviance: 13817  on 5234  degrees of freedom
## AIC: 16905
##
## Number of Fisher Scoring iterations: 6
```

Graphical representations to check the distribution of the residuals





The plots confirm a bit of heteroscedasticity because there is a pattern and points are not randomly spread: residual variability decreases as the fitted values increase. In a count model it's quite normal to have wider residuals where the fitted values are lower or higher because variance increases with the mean.

Thanks for your attention!

Health Data Science

Step by Passi