Analisi di sopravvivenza per il turnover aziendale

Mariangela Tafuri, Vincenzo Picarelli, Paolo Simari

Febbraio 2022

Scopo di questo report è quello di mostrare come l'analisi di sopravvivenza può aiutare le aziende a modellare e prevedere il turnover e l'uscita dei dipendenti. Segue una metodologia standard di scienza dei dati, con preparazione dei dati, analisi esplorativa, costruzione di modelli, valutazione e conclusioni per prevedere e descrivere il turnover dei dipendenti.

Parole chiave: *analisi di sopravvivenza, turnover, Kaplan-Meier, modello di regressione di Cox, R.*

Il caso studio

L'analisi di sopravvivenza viene utilizzata quando si intende analizzare un fenomeno in relazione a un periodo di tempo o analizzare il tempo trascorso tra un evento iniziale, in cui un soggetto o un oggetto entra in un particolare stato e un evento finale, che modifica questo stato.

In questo lavoro vengono discusse le tecniche di analisi della sopravvivenza per cercare di prevedere e modellare il fatturato all'interno delle aziende. Dopotutto, investire nella riduzione del turnover dei dipendenti, utilizzando la scienza dei dati per comprenderne le cause e le conseguenze, può comportare notevoli risparmi sui costi per i datori di lavoro.

Si è scelto di esaminare il caso studio di *Employee Turnover* (https://www.kaggle.com/davinwijaya/employee-turnover), che ci porta ad esaminare diversi fenomeni che potrebbero influenzare il turnover di un dipendente da un'azienda.

Analisi preliminare

Caricamento delle librerie, dati e controllo dei valori mancanti:

```
library(dplyr)
library(readr)
library(ggplot2)
library(gridExtra)
turnover <- read_csv("turnover.csv")
```

```
sum(is.na(turnover))
```

```
## [1] 0
```

Il comando riporta un dataset pulito di valori mancanti, tuttavia da ispezione preliminare, si nota che la variabile discreta *age* contiene valori anomali, continui che non si possono trattare. I valori in questione sono definiti in .cvs come "30.40033257". Pertanto vengono individuati come NA:

```
turnover$age<- ifelse((turnover$age %% 1)*10 > 0, NA, turnover$age)
sum(is.na(turnover$age))
```

```
## [1] 9
```

In questo modo, si identificano 9 NA che rappresentano il 0.7% del totale delle osservazioni e si decide di non trattarli, eliminandoli dalle analisi successive.

```
turnover<-na.omit(turnover)
dim(turnover)
```

```
## [1] 1120   16
```

Il dataset di riferimento è composto da 1120 osservazioni per 16 variabili, di cui 7 numeriche e 9 categoriche e si presenta nel seguente modo:

```
head(turnover)
```

```
## # A tibble: 6 x 16
##    stag event gender   age industry        profession traffic  coach head_gender
##   <dbl> <dbl> <chr>  <dbl> <chr>           <chr>      <chr>    <chr> <chr>
## 1  7.03     1 m         35 Banks           HR         rabrecN~ no    f
## 2 23.0      1 m         33 Banks           HR         empjs    no    m
## 3 15.9      1 f         35 PowerGeneration HR         rabrecN~ no    m
## 4 15.9      1 f         35 PowerGeneration HR         rabrecN~ no    m
## 5  8.41     1 m         32 Retail          Commercial youjs    yes   f
## 6  8.97     1 f         42 manufacture     HR         empjs    yes   m
## # ... with 7 more variables: greywage <chr>, way <chr>, extraversion <dbl>,
## #   independ <dbl>, selfcontrol <dbl>, anxiety <dbl>, novator <dbl>
```

La variabile *stag* rappresenta il tempo intercorso tra la data di assunzione e le dimissioni del dipendente espresso in mesi:

```
summary(turnover$stag)
```

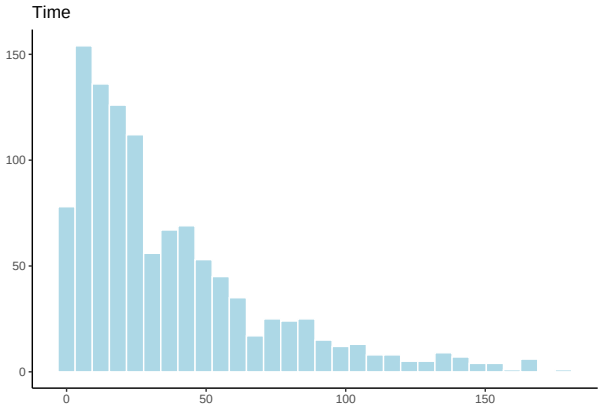
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.3942  11.7207  24.3121  36.6155  51.3183 179.4497
```

stag è una variabile continua che va da 0.4 a 179.4. Per questo studio, si considerano i mesi in valori interi e i valori minori di 1, vengono arrotondati all'unità, in questo modo:

```
turnover$stag<- round(ifelse(turnover$stag < 1, 1, turnover$stag),0)
summary(turnover$stag)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   12.00   24.00   36.62   51.00  179.00
```

```
ggplot(turnover, aes(x = stag)) +
  geom_histogram(fill = "lightblue",
                 color = "white") +
  labs(title="Time",
       x = "",
       y = "") +
  theme_classic()
```



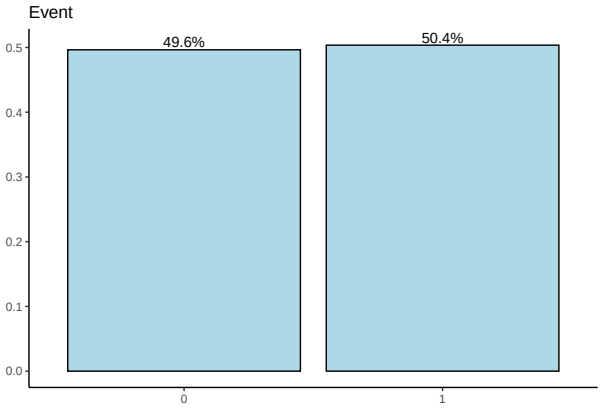
La variabile *event* assume valori di censura (0) e di avvenute dimissioni (1):

```
turnover %>% group_by(event) %>% summarise(n = n()) %>%
  mutate(freq = paste0(round(100 * n / sum(n), 1), "%"))
```

```
## # A tibble: 2 x 3
##   event     n freq
##   <dbl> <int> <chr>
## 1     0   556 49.6%
## 2     1   564 50.4%
```

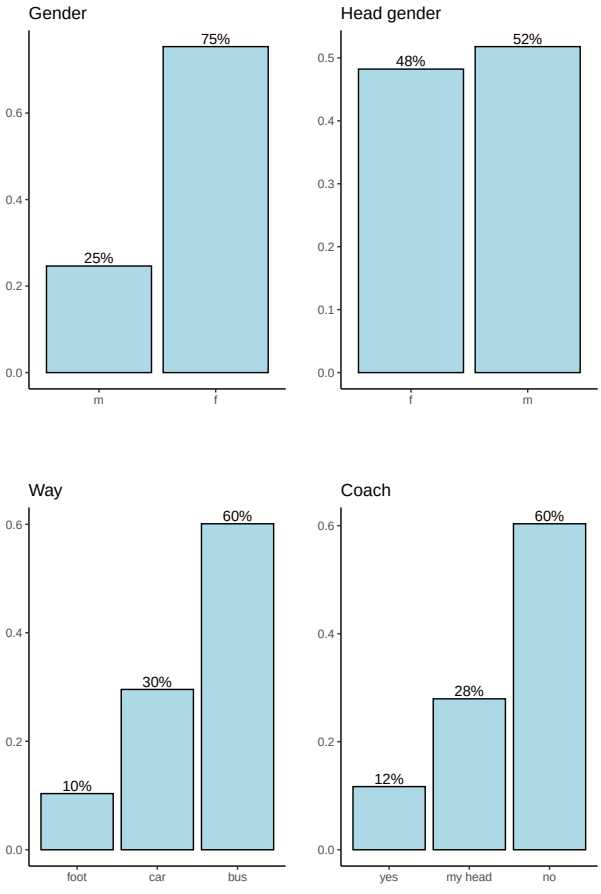
```
plot_event<- turnover %>%
  count(event) %>%
  mutate(pct = n / sum(n),
         pctlabel = paste0(round(pct*100, 1) , "%"))

ggplot(plot_event,
             aes(x = reorder(event, n),
                 y = pct)) +
  geom_bar(stat = "identity",
           fill = "lightblue",
           color = "black") +
  geom_text(aes(label = pctlabel),
            vjust=-0.25) +
  labs(x = "",
       y = "",
       title  = "Event") +
  theme_classic()
```

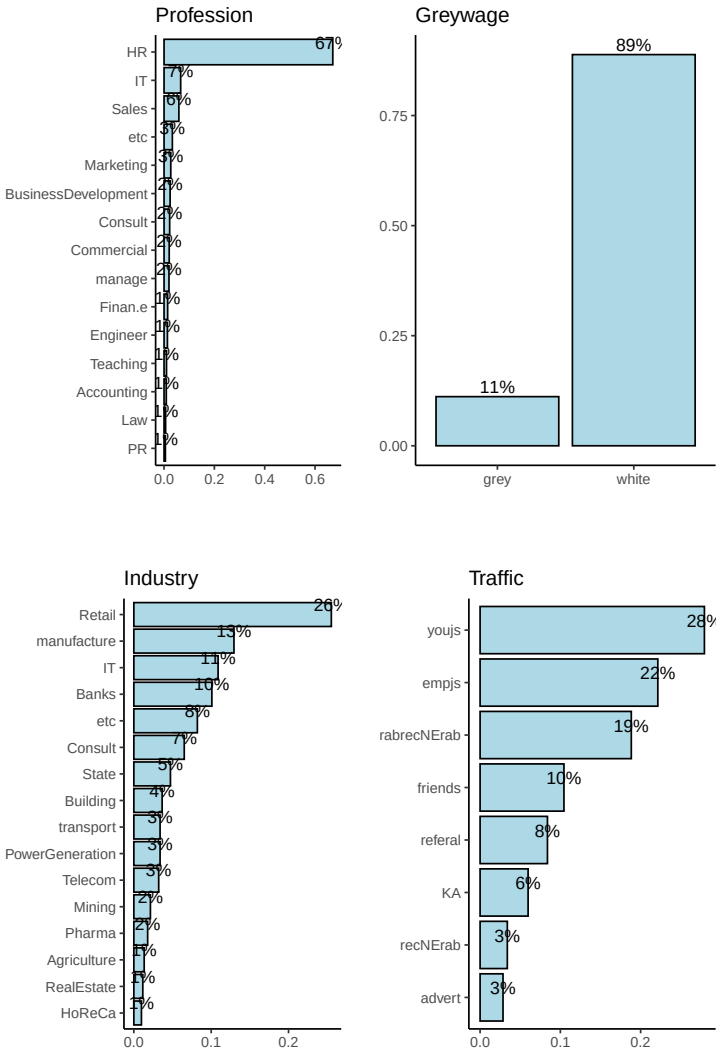


Le variabile categoriche sono:

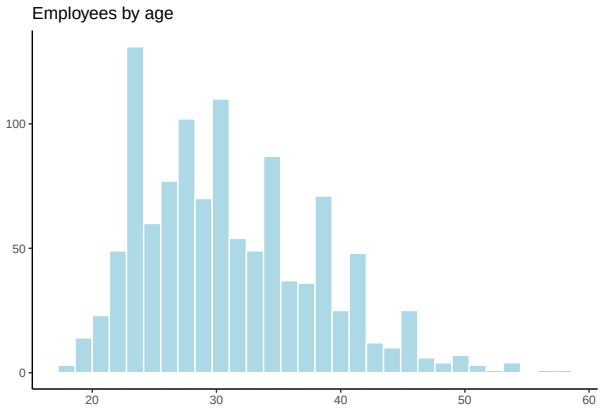
- il genere del dipendente (*Gender*), il modo in cui il dipendente raggiunge l'ufficio (*Way*), se il dipendente ha un mentore (*Coach*), il genere del capo (*Head gender*):

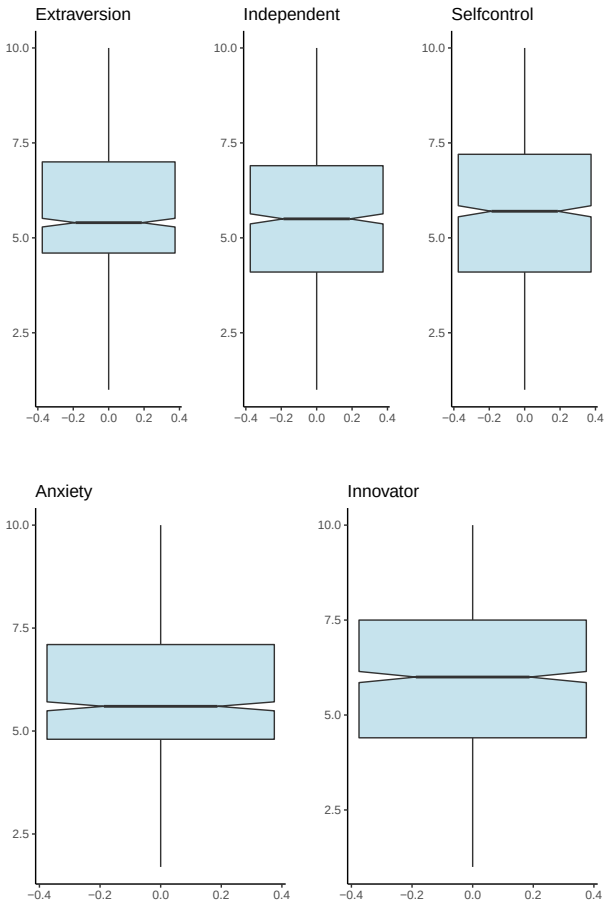


- l'ambito dell'azienda per cui lavora il dipendente (*industry*), la professione del dipendente (*profession*), da quale canale il dipendente è entrato in azienda (*traffic*), tipo di salario (*graywage*):



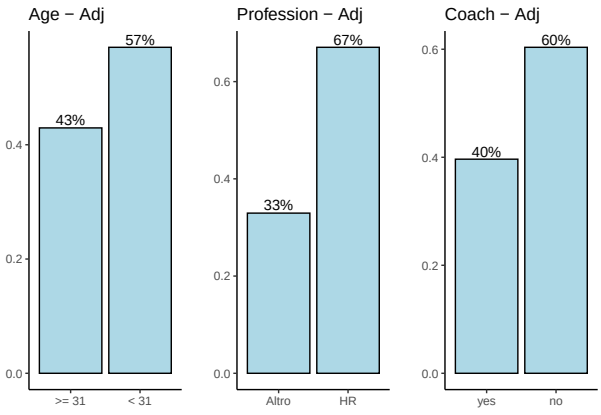
Mentre le variabili numeriche sono l'età del dipendente e le scale psicometriche che sintetizzano i livelli di estroversione, indipendenza, autocontrollo, ansia e innovazione:





Per analisi successive, si sono create nuove variabili a partire da queste:

- le dicotomizzazione della variabile età a livello del suo valore medio e delle variabili *profession* e *coach* per ridurre il numero di categorie:



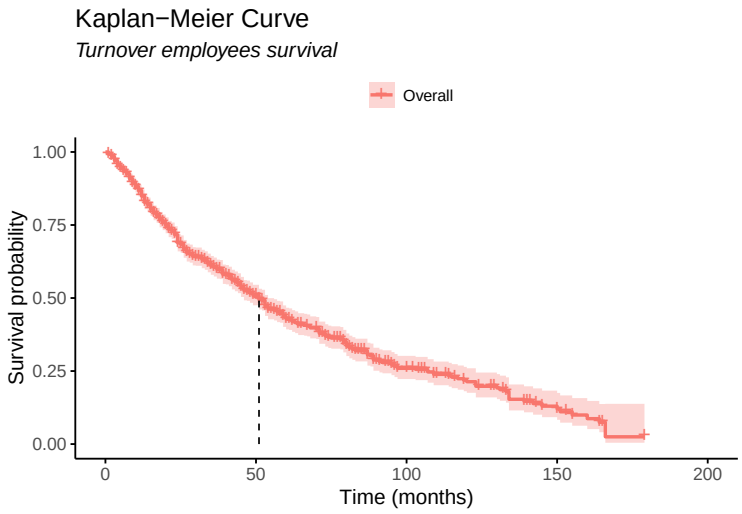
Analisi di sopravvivenza

Kaplan-Meier

```
library(survival)
library(survminer)
km<-survfit(Surv(stag, event)~1, data=turnover)
km
```

```
## Call: survfit(formula = Surv(stag, event) ~ 1, data = turnover)
##
##         n events median 0.95LCL 0.95UCL
## [1,] 1120    564     51      46      54
```

```
ggsurvplot(km,
           pval=TRUE,
           legend.labs=c("Overall"),
           legend.title="",
           title="Kaplan-Meier Curve",
           subtitle="Turnover employees survival",
           xlab = "Time (months)",
           surv.median.line = c("v"),
  ggtheme = theme_survminer(
             font.main = c(16),
             font.submain = c(13, "italic"),
             font.caption = c(13),
             font.x = c(13),
             font.y = c(13),
             font.tickslab = c(11)))
```



La mediana della probabilità di sopravvivenza si trova al mese 51 [IC95%: 46 - 54]: intorno a quel periodo il 50% dei dipendenti ancora non si è licenziato.

Volendo confrontare la probabilità di sopravvivenza in base all'età, si prende in considerazione la media come cut-point per discriminare la componente che ha meno di 31 anni, rispetto a chi ha più di 31 anni.

```
km_eta<-survfit(Surv(stag, event)~eta, data=turnover)
km_eta
```

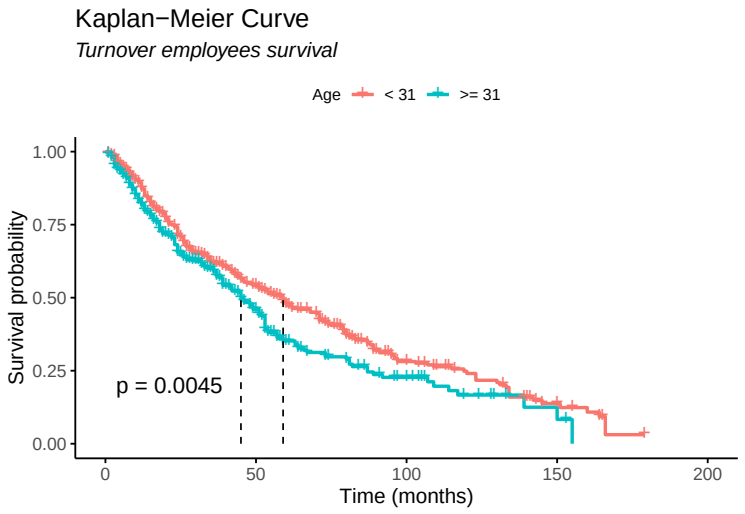
```
## Call: survfit(formula = Surv(stag, event) ~ eta, data = turnover)
##
##             n events median 0.95LCL 0.95UCL
## eta=< 31  639    326     59      49      70
## eta=>= 31 481    238     45      39      52
```

Per il gruppo che ha meno di 31 anni, la mediana della probabilità di sopravvivenza si trova al mese 59 [IC95%: 49 - 70]; mentre per il gruppo che ha più di 31 anni, la mediana della probabilità di sopravvivenza si trova al mese 45 [IC95%: 39 - 52]: intorno a questi periodi il 50% dei dipendenti ancora non si è licenziato.

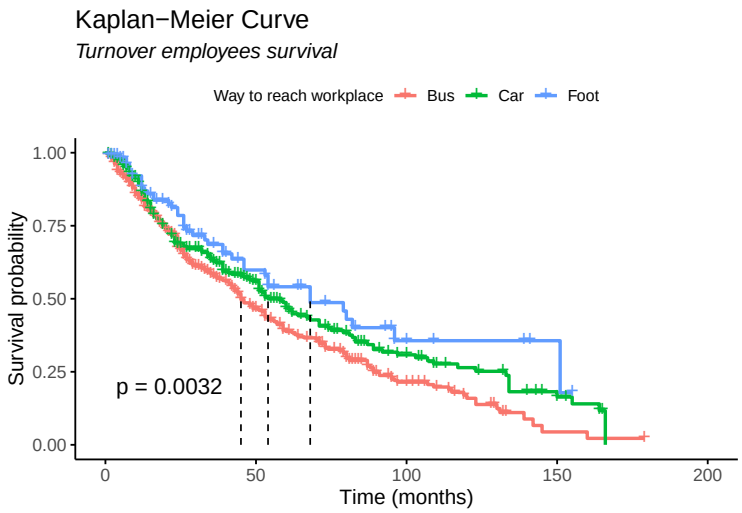
```
survdiff(Surv(stag, event)~eta, data=turnover)
```

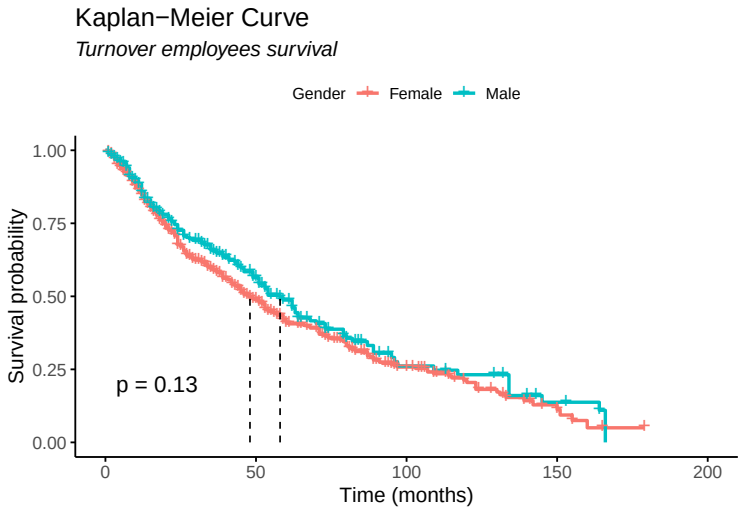
```
## Call:
## survdiff(formula = Surv(stag, event) ~ eta, data = turnover)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## eta=< 31  639      326      358      2.83      8.06
## eta=>= 31 481      238      206      4.92      8.06
##
##  Chisq= 8.1  on 1 degrees of freedom, p= 0.005
```

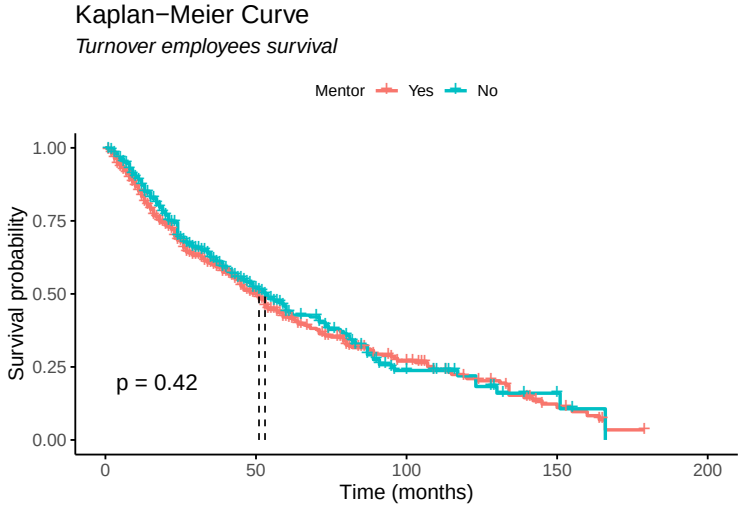
Con il Log-rank test ci si accerta che esiste una differenza tra probabilità di sopravvivenza tra i due gruppi. In questo caso si rifiuta l'ipotesi di uguaglianza tra le sopravvivenze dei diversi gruppi con un *p-value* = 0.005, riportato anche nel grafico seguente:

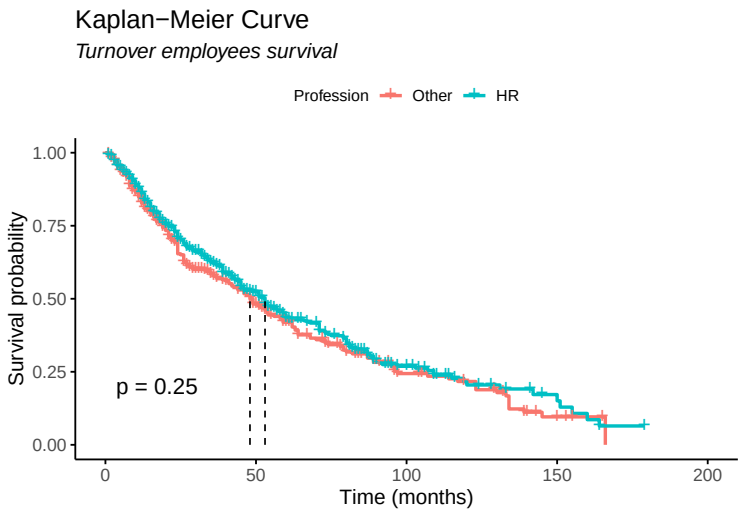


Si costruiscono altre curve di Kaplan-Meier, il cui valore *p-value* del Log-rank test è visibile sul grafico:


Kaplan−Meier Curve
Turnover employees survival
Way to reach workplace
Bus
Car
Foot
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.0032
0
50
100
150
200
Time (months)


Kaplan−Meier Curve
Turnover employees survival
Gender
Female
Male
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.13
0
50
100
150
200
Time (months)


Kaplan−Meier Curve
Turnover employees survival
Mentor
Yes
No
Survival probability
1.00
0.75
0.50
0.25
0.00
p = 0.42
0
50
100
150
200
Time (months)



Cox Regression

Si specifica un modello di Cox a rischi proporzionali con l'intento di stimare gli hazard ratio. Si sceglie di stabilire ad $\alpha = 0.05$ la significatività dei parametri. In un primo modello di prova si inserisco le variabili psicometriche, la variabile età e quella relativa al modo in cui i dipendenti raggiungono il posto di lavoro.

```
cox_mod_ampio<-coxph(Surv(stag, event) ~ age + way + extraversion
            + anxiety + selfcontrol + independ + age*way
            , data=turnover)
cox_mod_ampio
```

```
## Call:
## coxph(formula = Surv(stag, event) ~ age + way + extraversion +
##     anxiety + selfcontrol + independ + age * way, data = turnover)
##
##                   coef exp(coef)  se(coef)      z      p
## age           0.019052  1.019235  0.007539  2.527 0.0115
## waycar       -1.083845  0.338292  0.460716 -2.353 0.0186
## wayfoot       0.173595  1.189574  0.740604  0.234 0.8147
## extraversion  0.028323  1.028727  0.033086  0.856 0.3920
## anxiety      -0.043911  0.957039  0.031763 -1.382 0.1668
## selfcontrol  -0.049241  0.951952  0.030068 -1.638 0.1015
## independ     -0.004764  0.995248  0.033389 -0.143 0.8866
## age:waycar    0.027158  1.027531  0.014258  1.905 0.0568
## age:wayfoot  -0.021290  0.978935  0.024759 -0.860 0.3899
##
## Likelihood ratio test=42.86  on 9 df, p=2.288e-06
## n= 1120, number of events= 564
```

In generale il modello non raggiunge la significatività stabilita, tranne per la variabile età e la categoria per chi usa l'auto. Per tanto, si procede ad una pulizia del modello.

Si specifica, ora, un modello ridotto con le variabili che raggiungono il livello di significatività del 5%. Anche gli intervalli di confidenza confermano la significatività del modello, in quanto non comprendono il valore 1:

```
cox_mod_ristretto<-coxph(Surv(stag, event) ~ age+way+selfcontrol
            , data=turnover)
summary(cox_mod_ristretto)
```

```
## Call:
## coxph(formula = Surv(stag, event) ~ age + way + selfcontrol,
##     data = turnover)
##
##   n= 1120, number of events= 564
##
##                   coef exp(coef)  se(coef)      z Pr(>|z|)
## age          0.022682  1.022941  0.006122  3.705 0.000212 ***
## waycar      -0.260167  0.770922  0.094199 -2.762 0.005747 **
## wayfoot     -0.463295  0.629207  0.163734 -2.830 0.004661 **
## selfcontrol -0.061048  0.940778  0.021099 -2.893 0.003811 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## age            1.0229     0.9776    1.0107    1.0353
## waycar         0.7709     1.2971    0.6410    0.9272
## wayfoot        0.6292     1.5893    0.4565    0.8673
## selfcontrol    0.9408     1.0629    0.9027    0.9805
##
## Concordance= 0.571  (se = 0.014 )
## Likelihood ratio test= 33.07  on 4 df,   p=1e-06
## Wald test            = 33.03  on 4 df,   p=1e-06
## Score (logrank) test = 33.21  on 4 df,   p=1e-06
```

Inoltre, è verificata l'ipotesi di proporzionalità degli hazard con il seguente test:

```
cox.zph(cox_mod_ristretto)
```

```
##             chisq df    p
## age         0.398  1 0.53
## way         0.820  2 0.66
## selfcontrol 1.292  1 0.26
## GLOBAL      2.382  4 0.67
```

Il test implementato consente di verificare se esistono differenze significative tra le funzioni di sopravvivenza che vengono determinate sulle modalità di uno stesso carattere. Per ciascuna variabile, come anche per il modello nel suo complesso, non viene rigettata l'ipotesi di proporzionalità degli hazard.

Infine, i livelli di AIC per il primo e il secondo modello dimostrano che per parsimonia e significatività dei parametri, il modello ristretto è migliore perché registra il livello di AIC più basso.

```
library(lmtest)
AIC(cox_mod_ampio) # modello ampio
```

```
## [1] 6827.445
```

```r
AIC(cox_mod_ristretto) # modello significativo per alpha < 5%
```

```
## [1] 6827.23
```

Inoltre, si utilizza il test di Wald per i modelli *nested*: con un *p-value* di 0.08 non c'è evidenza sul miglioramento del modello ampio, per questo il modello ristretto è migliore.

```r
waldtest(cox_mod_ristretto, cox_mod_ampio)
```

```
## Wald test
##
## Model 1: Surv(stag, event) ~ age + way + selfcontrol
## Model 2: Surv(stag, event) ~ age + way + extraversion + anxiety + selfcontrol +
##     independ + age * way
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    560
## 2    555  5 9.7552    0.08248 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si analizzano i risultati:

- Volendo focalizzarsi sulla variabile età: a parità degli altri fattori, il rischio di licenziarsi aumenta del 2.3% per ogni anno di età in più. Per esempio, confrontando il rischio di licenziarsi relativo a un dipendente di 20 anni rispetto ad uno di 36, questo è di circa 1.4 volte superiore ad un impiegato di 20.

```r
exp(0.022682*16) # exp(coef di età * 16 anni)
```

```
## [1] 1.437509
```

- In più si può affermare che, a parità di tutti gli altri fattori, il rischio di licenziarsi si riduce del 6% con l'aumento unitario del livello di autocontrollo. Confrontando il rischio di licenziarsi relativo a un dipendente che ha un valore di *selfcontrol* minimo rispetto ad uno che registra un livello massimo, questo si riduce di 0.42.

```r
1 - exp(-0.061048*9)
```

```
## [1] 0.4227224
```

- Inoltre, si può constatare che il rischio di licenziarsi di chi va in ufficio con l'auto si riduce del 26% rispetto a chi va in bus; e il rischio di licenziarsi di chi va in ufficio a piedi si riduce del 37% rispetto a chi va in bus.

Conclusioni

Lo scopo del report era quello di mostrare come l'analisi di sopravvivenza, può aiutare le aziende a modellare e prevedere l'uscita dei dipendenti, siccome si tratta di un fenomeno che grava sul fatturato aziendale.

In tale sede, non si riesce a prevedere di quanto il turnover gravi sul fatturato aziendale, tuttavia si sono compresi alcuni meccanismi che influenzano l'uscita dei dipendenti e agire su questi fattori può sicuramente avere effetti positivi sul fenomeno.

Ad esempio, si nota che usare il bus come mezzo per raggiungere il posto lavorativo fa aumentare la probabilità di turnover rispetto a chi usa l'auto o va a piedi. In questo caso, si può pensare di stabilire giornate di smart-working o bonus per i mezzi di trasporto per chi li utilizza.

Altri accorgimenti possono essere adottati anche sul fronte psicologico: stabilire giornate o webinair dedicati a come acquisire più autocontrollo può migliorare lo status psicologico dei dipendenti e quindi, indirettamente, anche lo status del fenomeno di turnover aziendale.