Statistical Models - Project Work

Master in Artificial Intelligence and Data Science a.a. 2024/2025

Marco Longo - Francesca Ricci - Maria Rotella

2025-03-15

ANALISI DI SOPRAVVIVENZA

Caso di studio: permanenza dei dipendenti in azienda

Indice

Introduzione

Obiettivo

Utilizzando le tecniche apprese nel corso *Statistical Models*, il presente studio applica l'**analisi di sopravvivenza** al tema della permanenza dei dipendenti in azienda (*employee churn*).
L'obiettivo è identificare il profilo dei dipendenti a maggior rischio di abbandono, analizzando i fattori che ne influenzano *significativamente* il tempo all'evento, ossia la fine del rapporto di lavoro tra il dipendente e l'azienda.

Fonte dei dati

Il dataset utilizzato è un campione estratto dal dataset **hr-dataset** disponibile su *Kaggle.com*. Esso consiste in un file CSV con 500 osservazioni, ciascuna riferita ad un dipendente.

Allegati

Si allegano alla presente:

- il file Rmarkdown con cui è stata generata la relazione finale
- il dataset CSV utilizzato

Analisi di Sopravvivenza

Dataset

Il dataset contiene 500 osservazioni, una per ogni dipendente, ed include 10 variabili di tipo numerico e di tipo categoriale.

Per l'analisi di sopravvivenza, le due variabili chiave sono:

- *time_spend_company*: indica gli **anni** trascorsi in azienda (*follow up*);
- *left*: variabile binaria che rappresenta l'**evento** oggetto di studio, indicando l'uscita del dipendente (1 = uscita, 0 = permanenza). La tipologia di uscita (dimissioni, licenziamento, scadenza o mancato rinnovo) non è specificata e non rientrerà nell'analisi.

Le altre covariate:

- *satisfaction_level*: indice di soddisfazione (decimale tra 0 ed 1)
- *last_evaluation*: ultima valutazione ricevuta (decimale tra 0 ed 1)
- *number_project*: numero di progetti in carico al dipendente
- *average_monthly_hours*: media delle ore mensili lavorate
- *work_accident*: incidenti sul lavoro (0 = no, 1 = sì)
- *promotion_last_5years*: promozioni negli ultimi 5 anni (0 = no, 1 = sì)
- *department*: reparto aziendale (es. marketing, support, sales, HR, IT)
- *salary*: livello salariale (low, medium, high)

Analisi preliminare

In questa sezione viene fatta un'analisi preliminare del dataset, che ne aiuta a comprendere la struttura sia attraverso descrizioni statistiche sia tramite rappresentazioni grafiche.

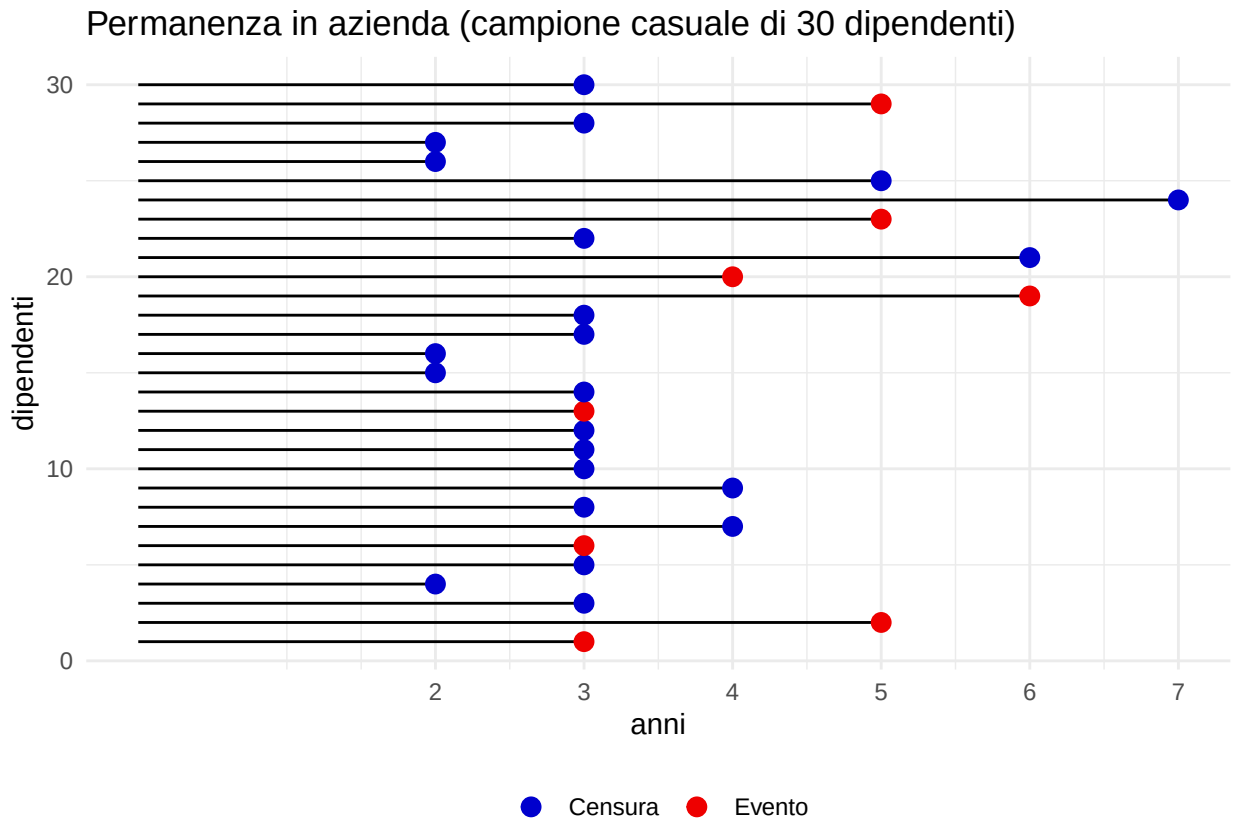
Evento e censura

La tabella riporta il numero di dipendenti che hanno lasciato l'azienda (evento) e quelli per cui l'evento non si è verificato durante il follow up (censura).

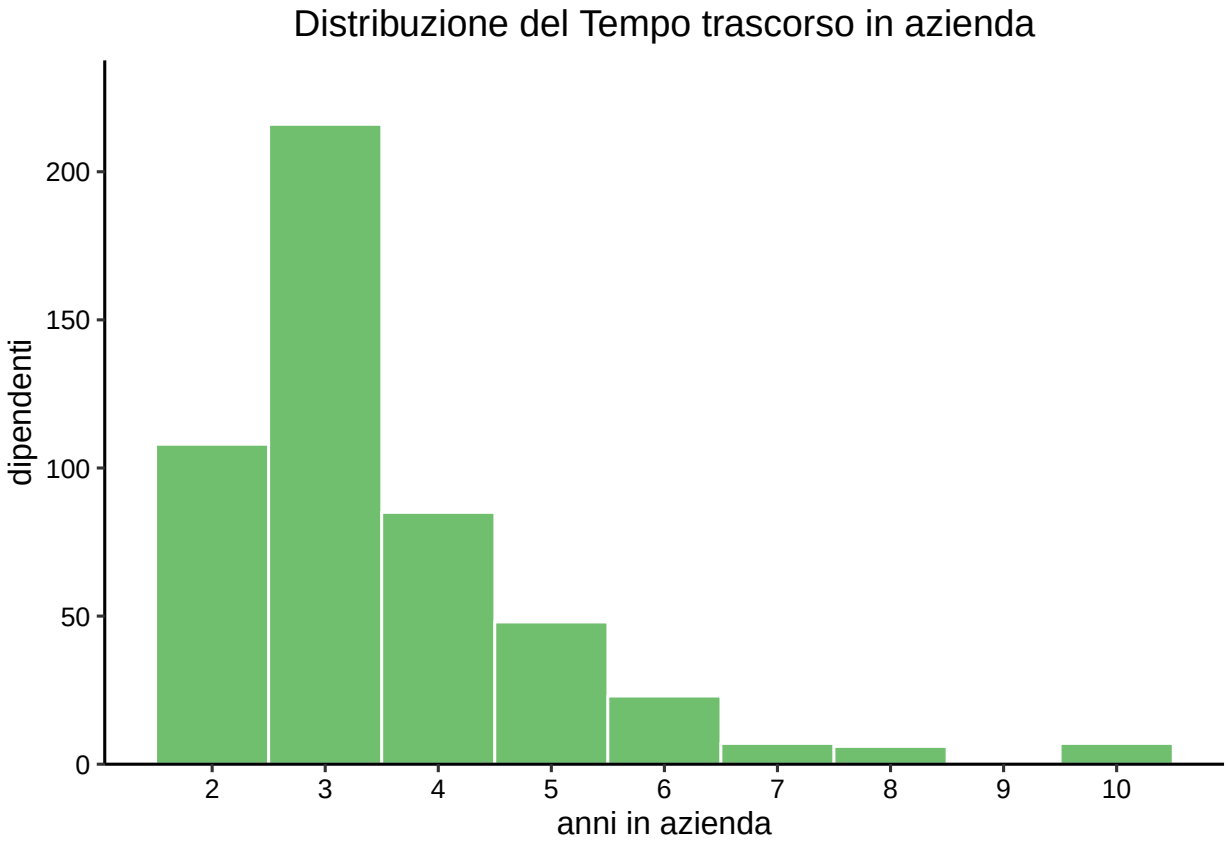
Si conta circa un quarto dei dipendenti per cui l'evento si manifesta.

	count	perc (%)
censura	380	76
evento	120	24

A scopo puramente illustrativo si visualizzano, nel grafico seguente, 30 dipendenti estratti casualmente dal dataframe. Si evidenziano i tempi differenti (riportati a *t0*) di permanenza in azienda, per quanto riguarda sia la censura che il manifestarsi dell'evento.

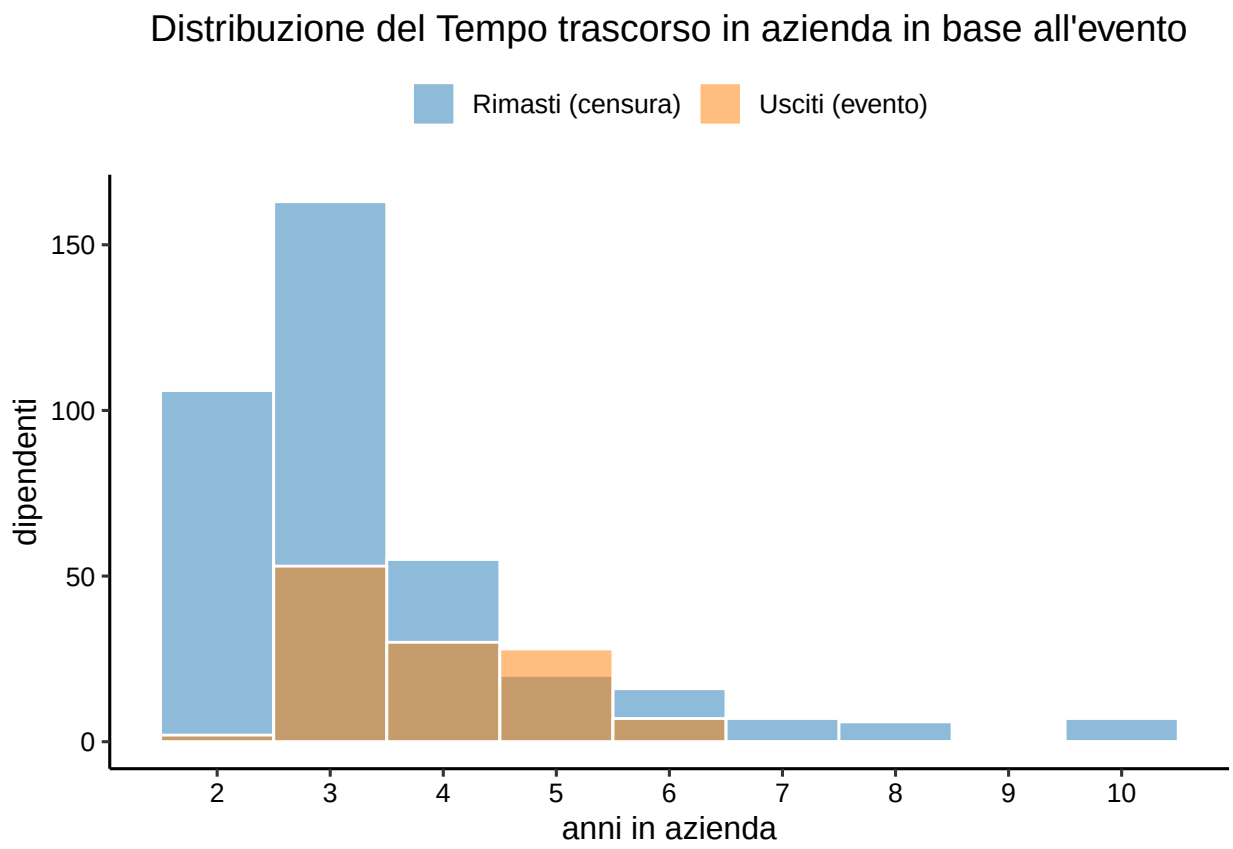


Tempo di permanenza



La media del tempo di permanenza per il campione osservato è di circa di 3 anni e mezzo.

count	min	q1	median	mean	q3	max	sd
500	2	3	3	3.498	4	10	1.466389



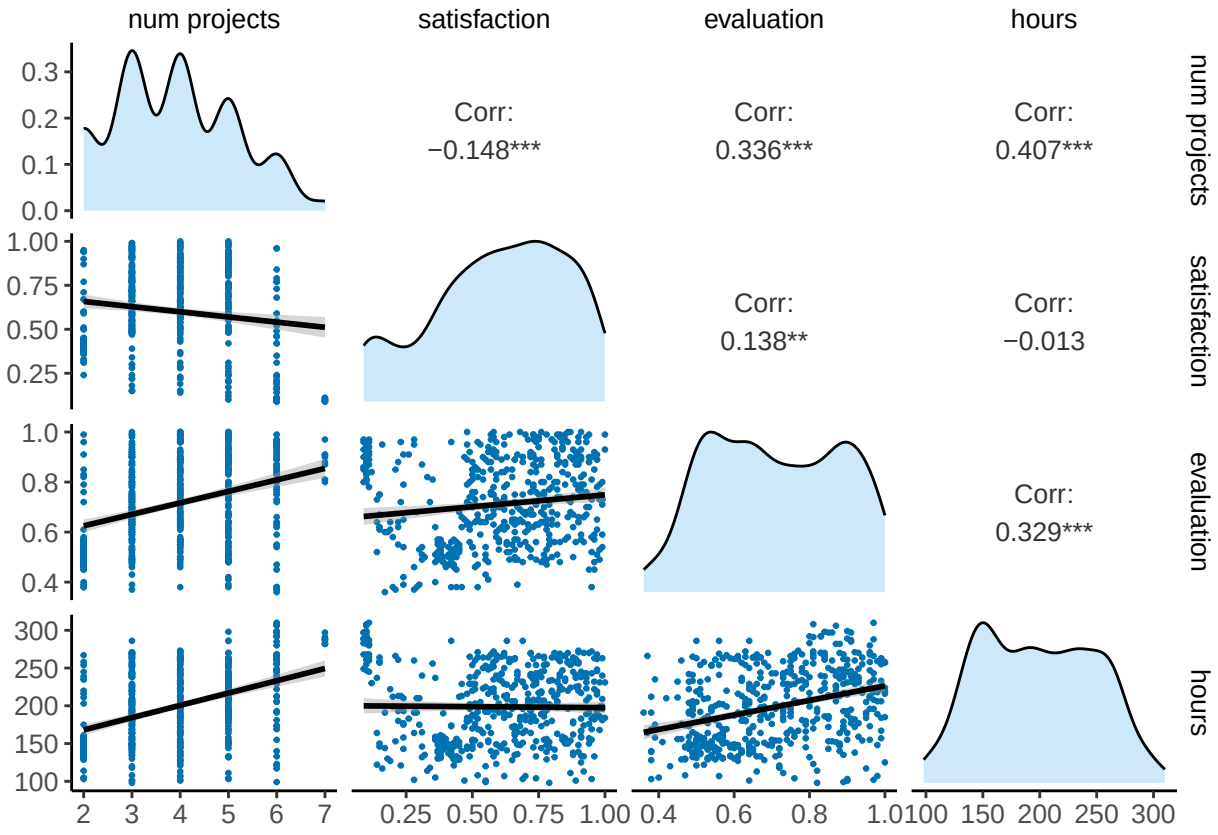
Di seguito la *summarise* che tiene conto dell'evento.
Il valore mediano del tempo trascorso in azienda di chi lascia è pari a 4 anni.

left	count	min	q1	median	mean	q3	max	sd
0	380	2	2	3	3.378947	4	10	1.5710370
1	120	2	3	4	3.875000	5	6	0.9835837

Covariate

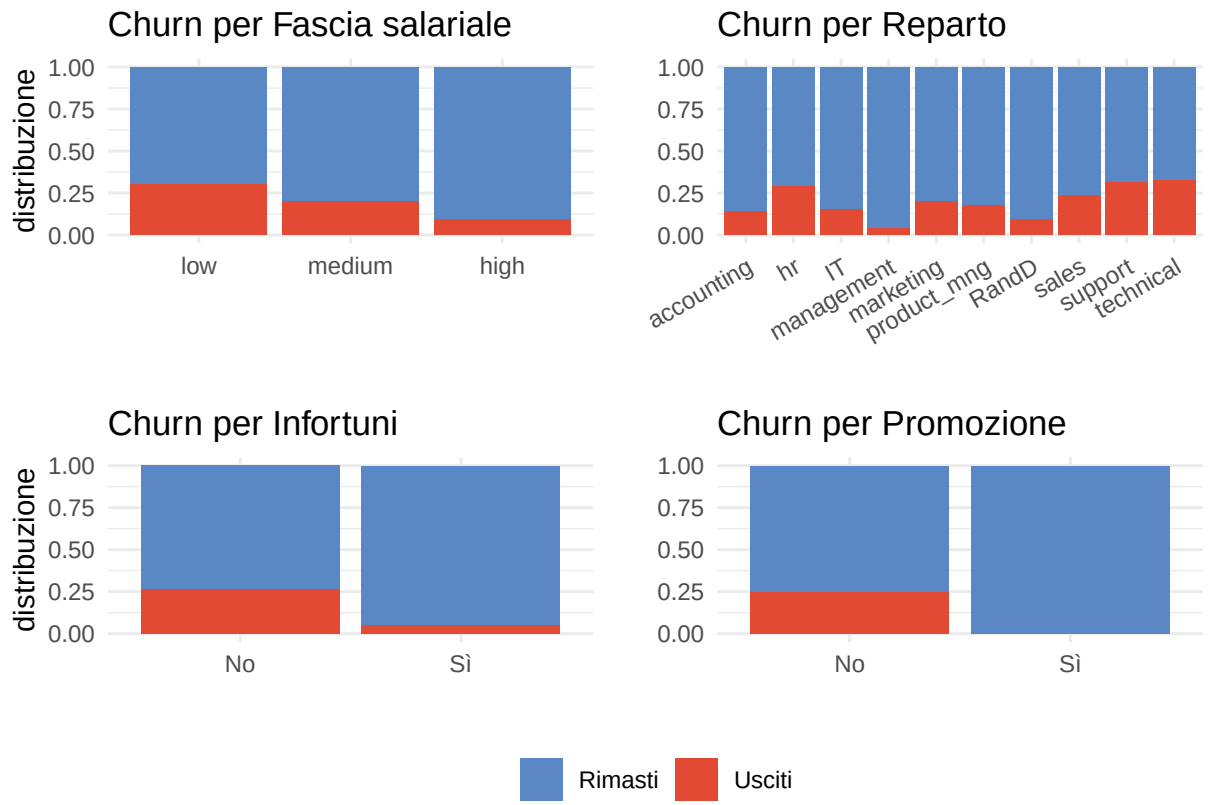
In questa sezione si effettua un'analisi preliminare delle covariate in relazione all'evento.

Il pair plot seguente evidenzia la correlazione tra le variabili numeriche del dataset.



Si evince una correlazione diretta tra il numero di progetti e le ore medie mensili lavorate, a loro volta correlate anche al livello di valutazione del dipendente: più ore lavora, più è alta la valutazione nei suoi confronti da parte dell'azienda.
Ma al crescere del carico di lavoro (ore e progetti), la soddisfazione percepita dal lavoratore decresce.

I grafici di seguito mostrano la distribuzione dell'**evento** in relazione ai valori delle covariate.

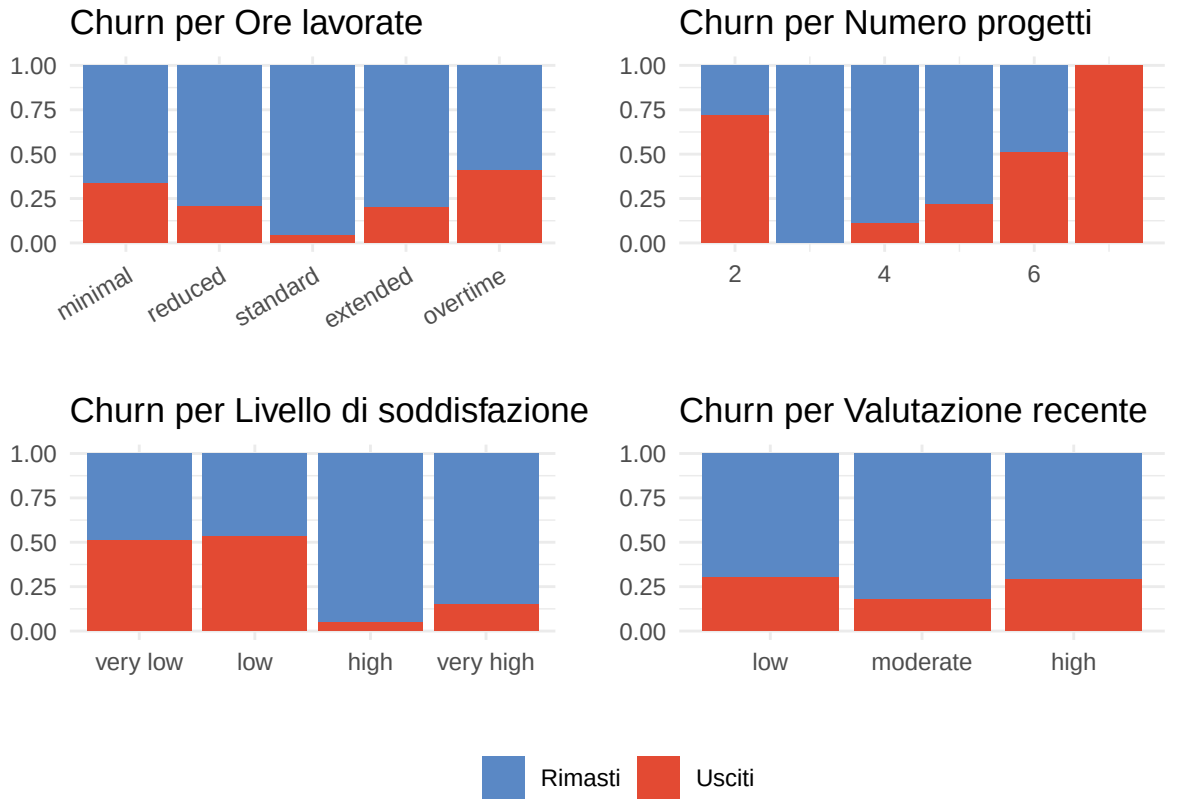


Come da aspettative, la **fascia salariale** sembra influire sulla permanenza in azienda: per stipendi bassi si registra un numero maggiore di abbandoni rispetto alle altre fasce.
Per alcuni dei **reparti** in cui si lavora, si notano delle differenze sulla distribuzione di abbandono: ad esempio di nota un minore numero di eventi per il *management* rispetto a chi lavora in *HR* o in reparti *tecnici*.
Si registrano meno abbandoni per chi ha avuto **infortuni** sul lavoro, probabilmente perché l'azienda potrebbe aver gestito bene quanto accaduto (mediante assicurazione o altre forme di attenzione e sostegno).
Come per gli stipendi, anche il fatto di aver ricevuto una **promozione** negli ultimi cinque anni sembra incidere sulla permanenza in azienda.

Sulle restanti covariate è necessario creare delle fasce per renderle categoriali. In particolare, sia per gli indici di soddisfazione e di valutazione (decimali da 0 a 1), sia per le ore mensili medie lavorate.

```
hr.df$hours_cat <- cut(hr.df$average_montly_hours,
                       breaks = quantile(hr.df$average_montly_hours,
                                         probs = seq(0, 1, length.out=6), na.rm = TRUE),
                       labels = c("minimal","reduced","standard","extended","overtime"),
                       include.lowest = TRUE)

hr.df$satisfaction_cat <- cut(hr.df$satisfaction_level, breaks = c(0,0.25,0.50,0.75,1),
                                  labels = c("very low", "low", "high", "very high"),
                                  include.lowest = TRUE)
hr.df$evaluation_cat <- cut(hr.df$last_evaluation, breaks = c(0.25, 0.5, 0.75, 1),
                                 labels = c("low", "moderate", "high"),
                                 include.lowest = TRUE)
```



Le **ore mensili** di lavoro presentano diversi livelli di chunk: il livello standard, corrispondente alla fascia centrale, risulta essere il più comunemente accettato dai dipendenti, mentre livelli *overtime* portano evidentemente il dipendente ad abbondonare l'azienda. Anche chi lavora poche ore ha un'alta probabilità che si manifesti l'evento, probabilmente perché si sente inutilizzato in base alle proprie aspettative o potenzialità.
Analogamente si può ragionare per numero di **progetti** assegnati: ad abbandonare l'azienda sono coloro che si sentono sovraccarichi (6-7 progetti contemporaneamente) o poco "utilizzati" (2 progetti), spesso sinonimo di poca fiducia.
Andamento come da aspettative anche per i livelli di **soddisfazione**: abbandona di più chi è insoddisfatto. La recente **valutazione** effettuata sul dipendente mostra abbandoni sulla fascia bassa, ma anche su quella alta, probabilmente perché i dipendenti più validi sono anche quelli più richiesti dal mercato.

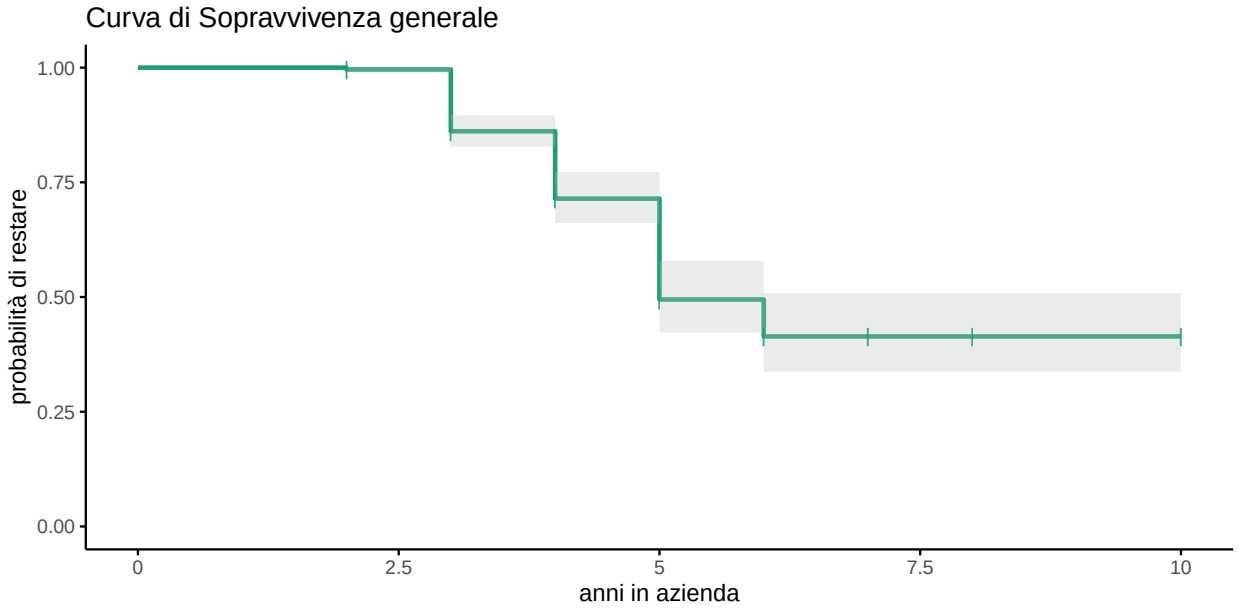
Stimatore di Kaplan Meier

L'analisi della sopravvivenza attraverso lo stimatore di Kaplan-Meier permette di osservare l'andamento della probabilità di permanenza dei dipendenti nel tempo, senza fare assunzioni specifiche sulla distribuzione dei tempi di sopravvivenza (stimatore non parametrico).
Si determina la configurazione dei dati attraverso la funzione *Surv* del pacchetto **survival**, considerando il tempo *time_spend_company* e l'evento *left*.
Lo stimatore KM è implementato dalla funzione *survfit*.

```
library(survival)
surv_object <- Surv(time = hr.df$time_spend_company, event = hr.df$left)
km_fit <- survfit(surv_object ~ 1, data = hr.df)
summary(km_fit)
```

```
## Call: survfit(formula = surv_object ~ 1, data = hr.df)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2    500       2    0.996 0.00282        0.990        1.000
##     3    392      53    0.861 0.01737        0.828        0.896
##     4    176      30    0.715 0.02835        0.661        0.772
##     5     91      28    0.495 0.03975        0.423        0.579
##     6     43       7    0.414 0.04340        0.337        0.509
```

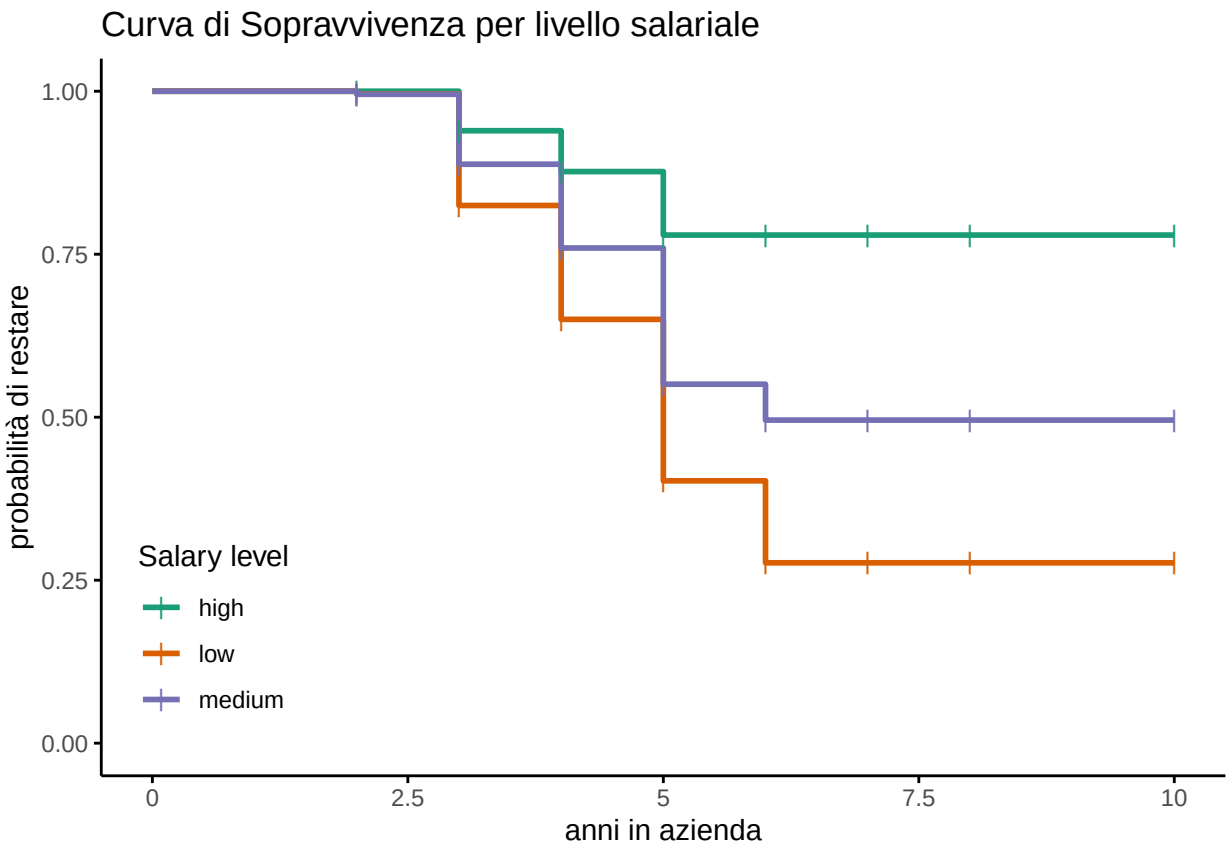
L'analisi evidenzia che circa il 70% dei dipendenti rimane in azienda dopo 4 anni. Tale percentuale scende drasticamente al quinto anno, sia per gli eventi che si manifestano ma anche, e soprattutto, perché il dipendente esce dall'osservazione. Il grafico seguente ne mostra l'andamento rispetto al tempo.



La curva di sopravvivenza mostra una progressiva riduzione della probabilità di restare in azienda, con un calo più marcato tra il terzo ed il quinto anno di lavoro.
Il valore degli intervalli di confidenza indica una certa variabilità nei dati, suggerendo che alcune categorie di dipendenti possono avere una certa incidenza sulla probabilità di abbandono rispetto alla media complessiva.

Curva di sopravvivenza per stipendio

La curva di sopravvivenza per livello salariale mostra chiaramente che i dipendenti con stipendio basso tendono ad abbandonare l'azienda prima di coloro che hanno uno stipendio medio o alto. La fascia *high* rimane significativamente sopra le altre, indicando una maggiore probabilità di rimanere in azienda nel lungo periodo.




```
## Call: survfit(formula = surv_object ~ salary, data = hr.df)
##
##                 salary=high
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     3     33       2    0.939  0.0415        0.861            1
##     4     15       1    0.877  0.0719        0.747            1
##     5      9       1    0.779  0.1119        0.588            1
##
##                 salary=low
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2    243       1    0.996 0.00411        0.988        1.000
##     3    192      33    0.825 0.02733        0.773        0.880
##     4     85      18    0.650 0.04242        0.572        0.739
##     5     42      16    0.402 0.05534        0.307        0.527
##     6     16       5    0.277 0.06018        0.181        0.424
##
##                 salary=medium
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2    214       1    0.995 0.00466        0.986        1.000
```

```
##     3    167      18    0.888 0.02424        0.842        0.937
##     4     76      11    0.760 0.04141        0.683        0.845
##     5     40      11    0.551 0.06145        0.442        0.685
##     6     20       2    0.496 0.06651        0.381        0.645
```

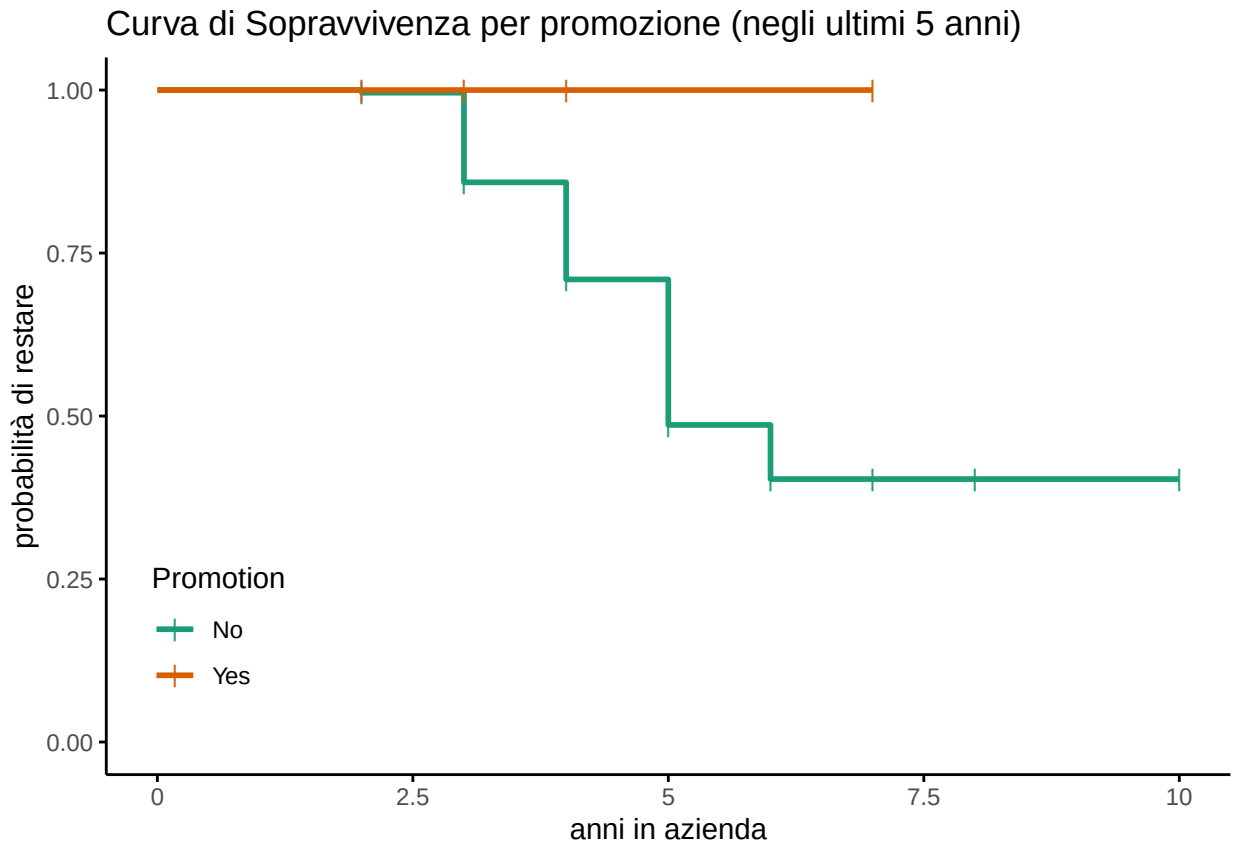
Come si legge dal *summary*, il **survival** (probabilità di sopravvivenza/permanenza) dopo 5 anni è pari al 78% per chi ha uno stipendio alto, al 55% per chi ha uno stipendio medio e al 40% per chi ha uno stipendio basso.

```
## Call:
## survdiff(formula = surv_object ~ hr.df$salary)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## hr.df$salary=high      43        4     11.1      4.54      6.13
## hr.df$salary=low      243       73     56.9      4.52     10.52
## hr.df$salary=medium 214       43     52.0      1.54      3.31
##
##  Chisq= 13  on 2 degrees of freedom, p= 0.002
```

Il test **Log-Rank**, implementato da **survdiff**, determina la significatività delle fasce rispetto alla permanenza in azienda. Si nota una significativa differenza tra il valore atteso e quello osservato, soprattutto per la fascia alta (4 eventi osservati su 11 attesi) e la fascia bassa (73 osservati su 57 attesi).
Il *p-value* rilevato, nettamente inferiori ai valori di soglia standard (0.05 o 0.01), permette di **rifiutare** l'ipotesi nulla H0, ipotesi secondo la quale il livello salariale non sia un fattore determinante nella permanenza in azienda.

Curva di sopravvivenza per promozione

Il dataset presenta una percentuale bassa (circa il 2%) di dipendenti che hanno ricevuto una promozione negli ultimi 5 anni. Per tale gruppo, nessuno ha manifestato l'evento e tutti sono stati censurati dopo al più 5 anni, come si evince anche dalla curva di sopravvivenza.

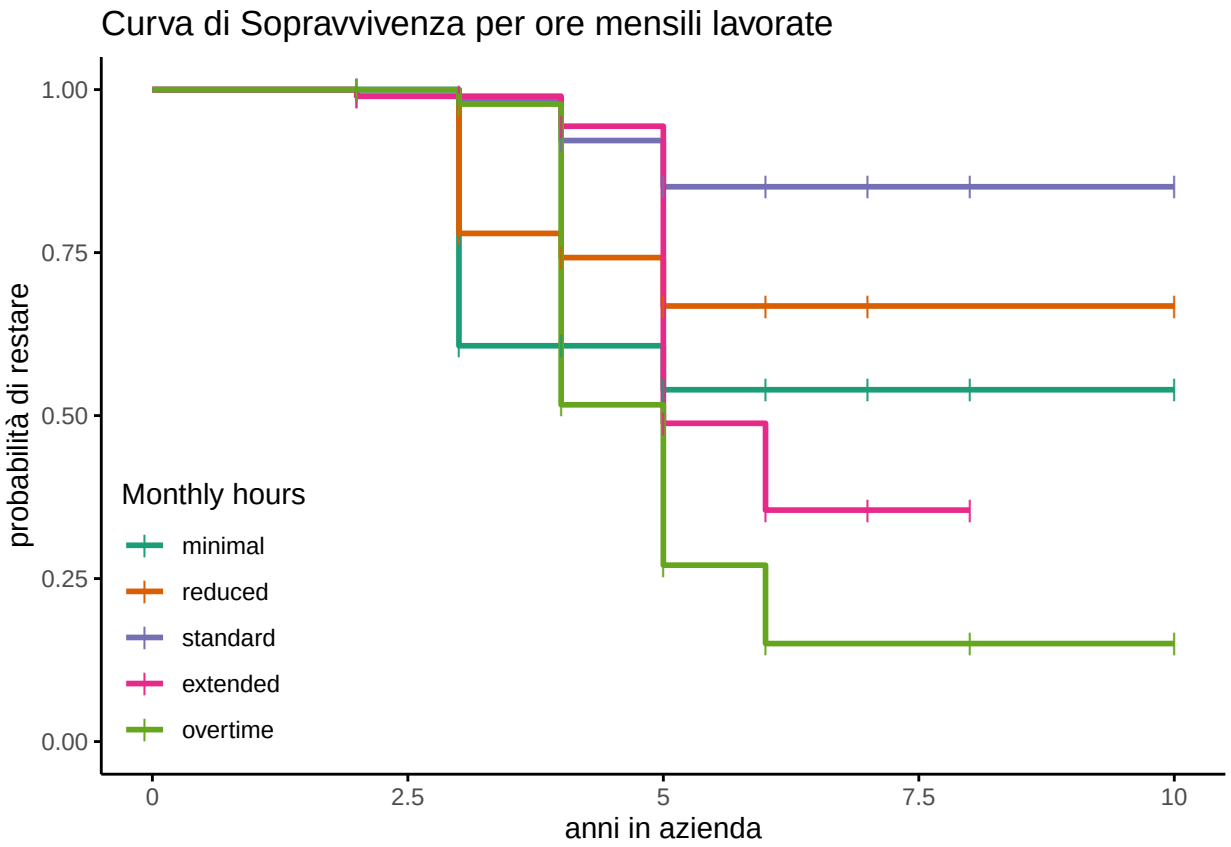


Il test Log-Rank non evidenzia una differenza tra valori attesi ed osservati tra i due gruppi. Il p-value infatti è superiore a 0.05, pertanto **non si rifiuta H0** secondo la quale i due valori di *promotion_last_5years* non incidono significativamente sulla sopravvivenza.

```
## Call:
## survdiff(formula = surv_object ~ hr.df$promotion_last_5years)
##
##                                 N Observed Expected (O-E)^2/E (O-E)^2/V
## hr.df$promotion_last_5years=0 491      120   117.43    0.0562      3.21
## hr.df$promotion_last_5years=1   9        0     2.57    2.5700      3.21
##
##  Chisq= 3.2  on 1 degrees of freedom, p= 0.07
```

Curva di sopravvivenza per ore mensili lavorate

La curva di sopravvivenza mostrata di seguito evidenzia una maggiore probabilità di permanenza per coloro che lavorano un numero di ore mensili *standard* (fascia media), mentre ad abbandonare l'azienda con maggiore probabilità sono coloro che lavorano *overtime*. Per questi ultimi, infatti, la probabilità di sopravvivenza scende a circa 0.50 dal quarto anno. Come evidenziato in precedenza, anche chi lavora poche ore ha una probabilità di lasciare l'azienda più alta rispetto alla fascia media.
Da notare anche l'accavallamento delle curve per i vari valori, il che potrebbe indicare una dipendenza dei rischi dal fattore tempo, come verrà approfondito successivamente con il modello di Cox.

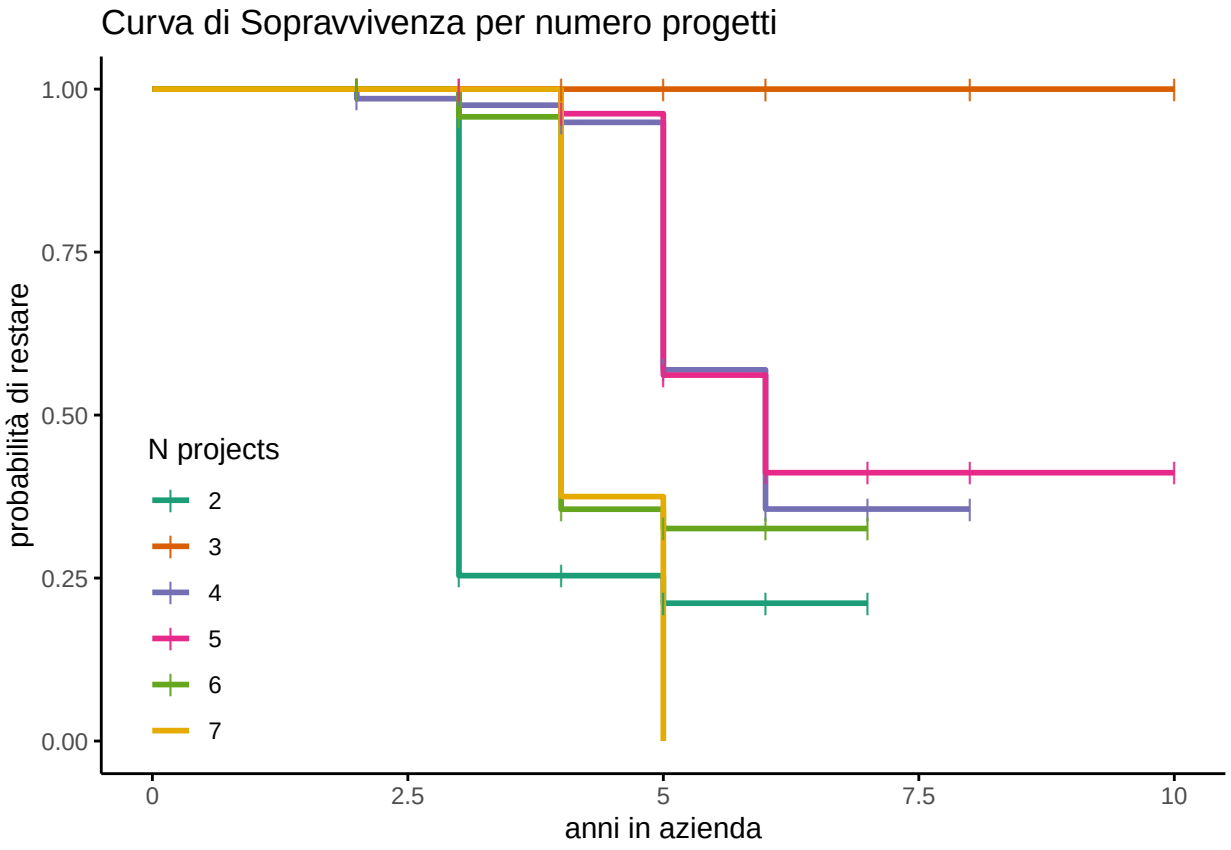


I risulati del test Log-Rank mostrati di seguito confermano che tale variabile risulta essere significativamente influente sulla permanenza in azienda: infatti i valori di *(O-E)^2/V* alti indicano una discrepanza rilevante tra valori attesi e valori osservati, ed il p-value estremamente piccolo conferma che l'ipotesi H0 viene rifiutata.

```
## Call:
## survdiff(formula = surv_object ~ hr.df$hours_cat)
##
##                              N Observed Expected (O-E)^2/E (O-E)^2/V
## hr.df$hours_cat=minimal    104       35     23.5     5.568     8.450
## hr.df$hours_cat=reduced     98       20     18.5     0.119     0.172
## hr.df$hours_cat=standard    99        4     20.6    13.412    19.652
## hr.df$hours_cat=extended    99       20     27.8     2.173     3.550
## hr.df$hours_cat=overtime   100       41     29.5     4.456     7.235
##
##  Chisq= 31.3  on 4 degrees of freedom, p= 3e-06
```

Curva di sopravvivenza per numero di progetti

La curva mostra come i dipendenti con un numero molto basso (2) o molto alto (6-7) di progetti abbiano una maggiore probabilità di lasciare l'azienda rispetto a coloro che lavorando su un numero di progetti intermedio, come già accennato nell'analisi preliminare.



Si noti come per il gruppo di dipendenti con il numero di progetti massimo il crollo della curva sia drastico. Dal *summary* di seguito si evince che degli 8 dipendenti che lavorano su 7 progetti, 5 abbandonano dopo 4 anni, gli 3 al quinto anno, portando a zero la *survival*.

```
## Call: survfit(formula = surv_object ~ number_project, data = hr.df)
##
##                 number_project=2
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     3     67      50    0.254  0.0532        0.168        0.383
##     5      6       1    0.211  0.0588        0.123        0.365
##
##                 number_project=3
##      time n.risk n.event survival std.err lower 95% CI upper 95% CI
##
##                 number_project=4
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2    136       2    0.985  0.0103        0.965        1.000
##     3     99       1    0.975  0.0142        0.948        1.000
##     4     37       1    0.949  0.0295        0.893        1.000
```

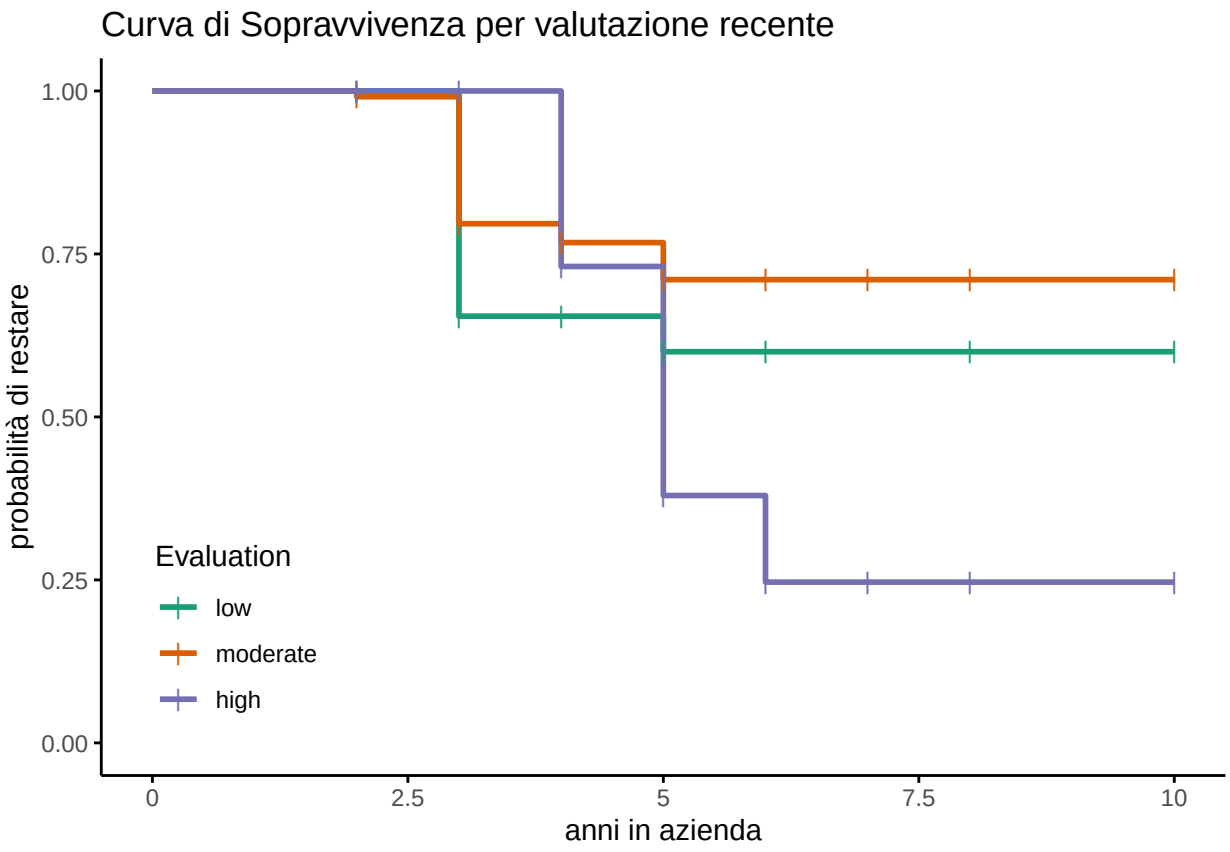
```
##     5     20       8    0.569  0.1054        0.396        0.819
##     6      8       3    0.356  0.1177        0.186        0.680
##
##                 number_project=5
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     4     53       2    0.962  0.0262        0.912        1.000
##     5     36      15    0.561  0.0805        0.424        0.744
##     6     15       4    0.412  0.0871        0.272        0.623
##
##                 number_project=6
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     3     47       2    0.957  0.0294        0.901         1.00
##     4     35      22    0.356  0.0790        0.230         0.55
##     5     12       1    0.326  0.0777        0.204         0.52
##
##                 number_project=7
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     4      8       5    0.375   0.171        0.153        0.917
##     5      3       3    0.000     NaN           NA           NA
```

Il test Log-Rank conferma che il numero di progetti assegnati influisce significativamente sulla durata della permanenza in azienda (p-value estremamente basso) e che la discrepanza tra valori osservati e valori attesi è molto alta, soprattutto per la fascia bassa (2 progetti).

```
## Call:
## survdiff(formula = surv_object ~ hr.df$number_project)
##
##                           N Observed Expected (O-E)^2/E (O-E)^2/V
## hr.df$number_project=2  71       51     12.9    112.85    156.03
## hr.df$number_project=3 139        0     25.8     25.79     39.97
## hr.df$number_project=4 136       15     27.7      5.82      9.21
## hr.df$number_project=5  97       21     33.2      4.49      7.93
## hr.df$number_project=6  49       25     17.0      3.74      5.36
## hr.df$number_project=7   8        8      3.4      6.22      7.95
##
##  Chisq= 193  on 5 degrees of freedom, p= <2e-16
```

Curva di sopravvivenza per valutazione recente

La valutazione delle performance presenta una curva di sopravvivenza con evidenti sovrapposizioni. I dipendenti con una valutazione alta mostrano una maggiore propensione all'abbandono, probabilmente perché associati ad un numero elevato di progetti, fattore che ha un impatto rilevante sull'abbandono, come descritto in precedenza.

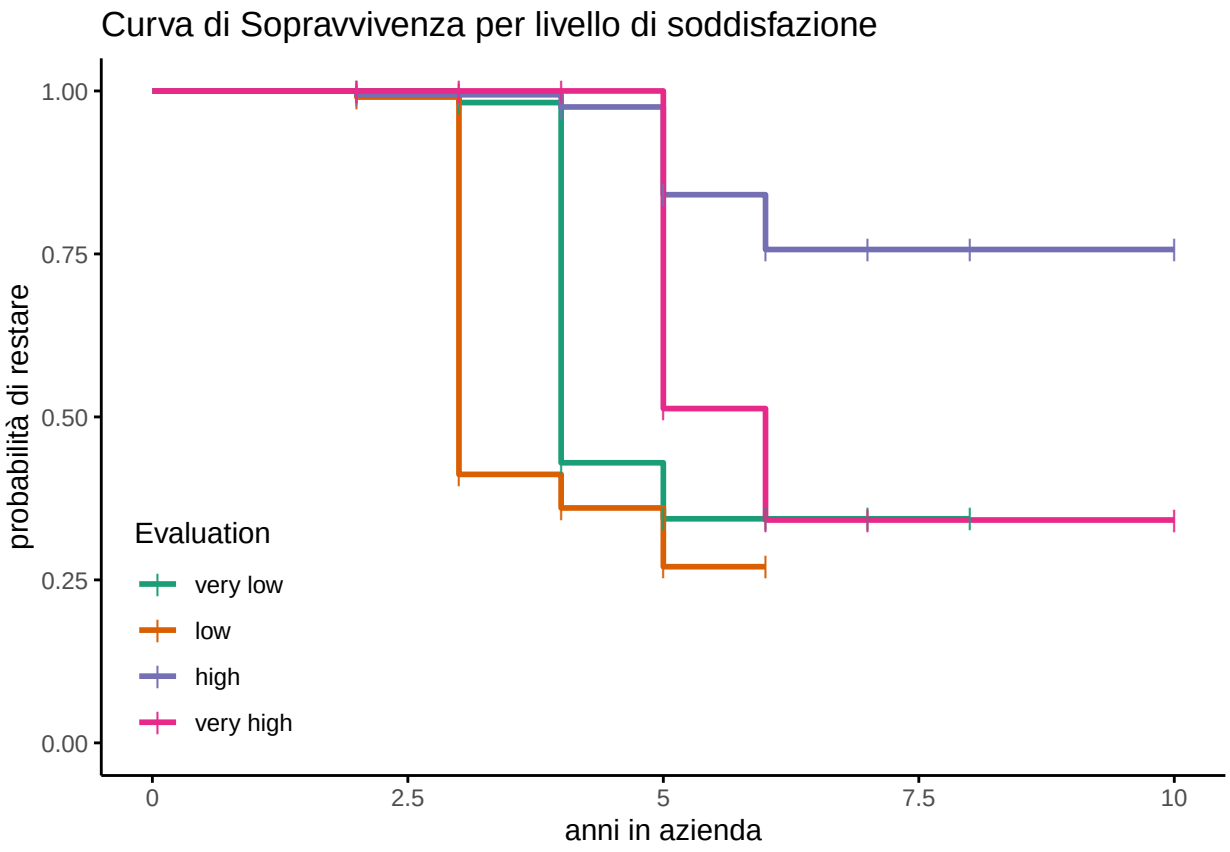


Il test Log-Rank indica che le differenze tra le categorie di valutazione non sono altamente significative. Il p-value infatti è > 0.05 ed anche in questo caso l'ipotesi H0 non si può rifiutare.

```
## Call:
## survdiff(formula = surv_object ~ hr.df$evaluation_cat)
##
##                                 N Observed Expected (O-E)^2/E (O-E)^2/V
## hr.df$evaluation_cat=low       66       20     15.3  1.47e+00  2.05e+00
## hr.df$evaluation_cat=moderate 226       40     44.7  5.03e-01  9.91e-01
## hr.df$evaluation_cat=high     208       60     60.0  2.18e-06  5.45e-06
##
##  Chisq= 2.4  on 2 degrees of freedom, p= 0.3
```

Curva di sopravvivenza per livello di soddisfazione

Anche per il livello di soddisfazione le curve di sopravvivenza si incrociano: dipendenti con livelli di soddisfazione bassi lasciano l'azienda molto più frequentemente rispetto a quelli con livelli medi. Ma anche chi dichiara un livello di soddisfazione molto alto dopo 5 anni presenta una probabilità dimezzata di rimanere in azienda.

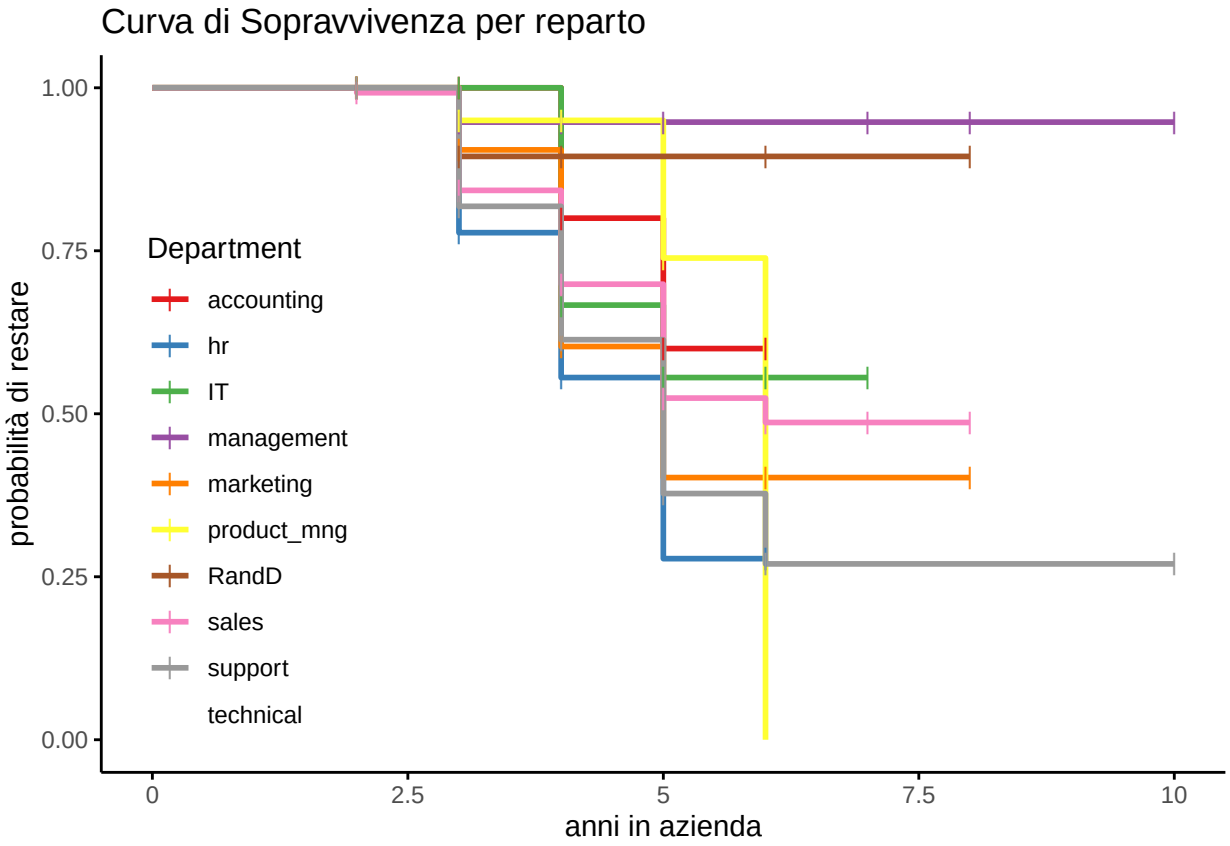


Il test Log-Rank conferma una differenza altamente significativa tra i gruppi, classificando il livello di soddisfazione del dipendente come uno dei fattori che influenzano la permanenza aziendale (p-value molto basso).

```
## Call:
## survdiff(formula = surv_object ~ hr.df$satisfaction_cat)
##
##                                   N Observed Expected (O-E)^2/E (O-E)^2/V
## hr.df$satisfaction_cat=very low   61       31     21.6      4.10      6.23
## hr.df$satisfaction_cat=low       107       57     18.0     84.72    123.75
## hr.df$satisfaction_cat=high      171        8     38.7     24.32     43.93
## hr.df$satisfaction_cat=very high 161       24     41.8      7.56     14.31
##
##  Chisq= 149  on 3 degrees of freedom, p= <2e-16
```

Curva di sopravvivenza per reparto

L'analisi per reparto mostra differenze marcate solo tra alcuni valori: in particolare si registra una maggiore sopravvivenza tra i dipendenti dei reparti *management* e *RandD*, e maggiori abbandoni per quelli del reparto *support* e *technical*. Per *product_mng* non si hanno osservazioni dopo i 6 anni, nonostante si registrino pochi eventi di abbandono rispetto alle censure.

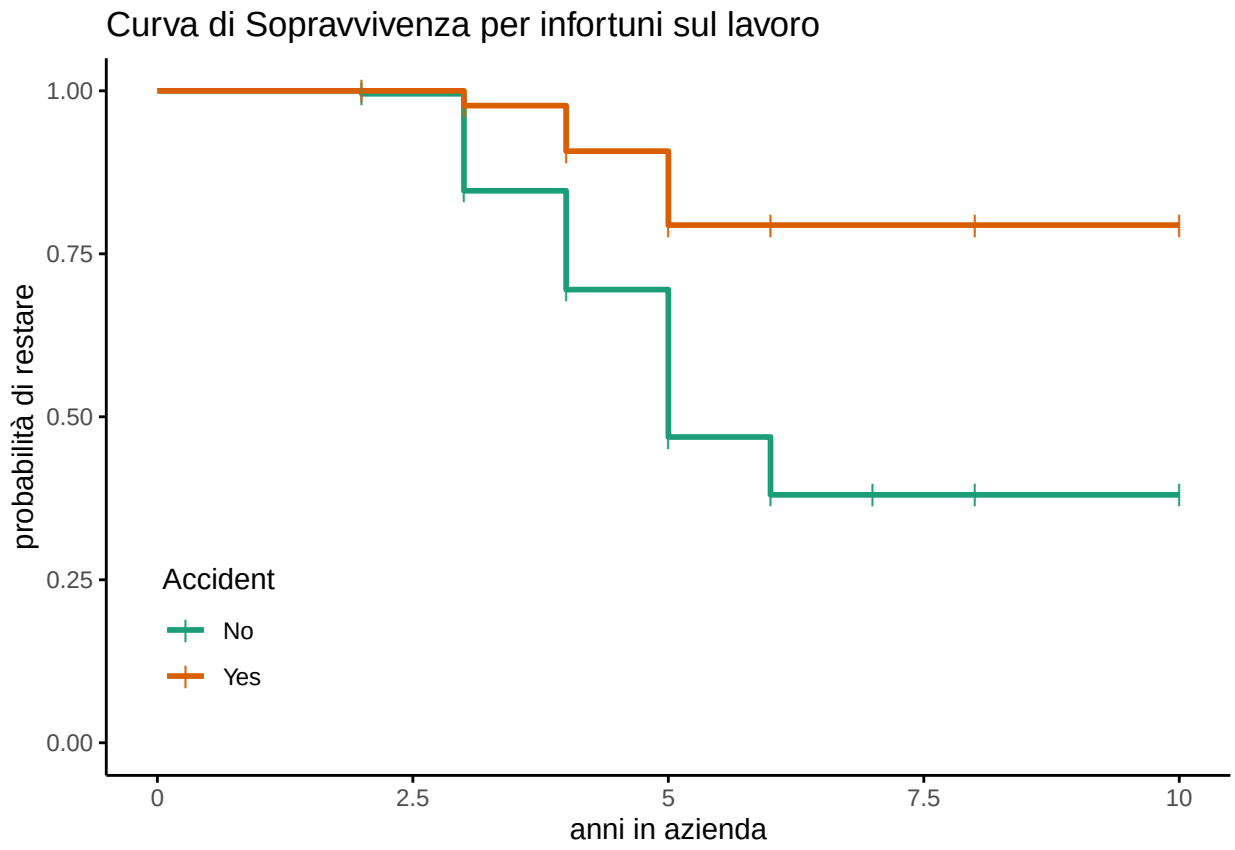


Il test Log-Rank mostra delle discrepanze tra valori attesi ed osservati solo per alcuni reparti. Nonostante il p-value non sia piccolo come per altri fattori, rimane comunque significativamente al di sotto della soglia standard, per cui si può rifiutare H0 e si può determinare il reparto come una variabile influente sull'abbandono dell'azienda.

```
## Call:
## survdiff(formula = surv_object ~ hr.df$department)
##
##                                N Observed Expected (O-E)^2/E (O-E)^2/V
## hr.df$department=accounting   21        3     5.48   1.12274    1.4350
## hr.df$department=hr           24        7     4.50   1.38743    1.7262
## hr.df$department=IT           33        5     7.86   1.04334    1.3603
## hr.df$department=management   24        1     8.32   6.43928    8.6198
## hr.df$department=marketing    30        6     5.74   0.01159    0.0146
## hr.df$department=product_mng  28        5     7.79   0.99683    1.3359
## hr.df$department=RandD        21        2     4.45   1.34578    1.6724
## hr.df$department=sales       132       31    31.51   0.00832    0.0138
## hr.df$department=support      86       27    19.86   2.56563    3.7306
## hr.df$department=technical   101       33    24.49   2.95935    4.5464
##
##  Chisq= 22  on 9 degrees of freedom, p= 0.009
```

Curva di sopravvivenza per infortunio sul lavoro

L'aver avuto un infortunio sul lavoro determina una minore probabilità di abbandono dell'azienda, probabilmente, come detto in fase preliminare, perché il dipendente riceve le opportune attenzioni (rimborso assicurativo?) ed è meno portato a lasciare l'azienda.



Anche il test Log-Rank mostra un valore osservato di eventi nettamente inferiore e quello atteso (3 vs 12) ed un p-value tale da rifiutare l'ipotesi H0. Pertanto può considerarsi un fattore significativamente influente rispetto al verificarsi dell'evento.

```
## Call:
## survdiff(formula = surv_object ~ hr.df$work_accident)
##
##                          N Observed Expected (O-E)^2/E (O-E)^2/V
## hr.df$work_accident=0 438      117      108     0.754      9.12
## hr.df$work_accident=1  62        3       12     6.770      9.12
##
##  Chisq= 9.1  on 1 degrees of freedom, p= 0.003
```

Modello di regressione di Cox

Nel presente studio si utilizza il **modello di rischi proporzionali di Cox** per stimare gli effetti delle variabili sul tempo di permanenza in azienda, senza fare assunzioni specifiche sulla forma della funzione di rischio di base. Questo modello presuppone che i rischi siano proporzionali nel tempo, consentendo così di valutare l'impatto delle variabili indipendenti sui tempi di sopravvivenza.
In questa sezione, è stata condotta la verifica della proporzionalità dei rischi per ciascuna covariata.

Stipendio

```
## Call:
## coxph(formula = surv_object ~ salary, data = hr.df)
##
##   n= 500, number of events= 120
##
##                coef exp(coef) se(coef)     z Pr(>|z|)
## salarylow    1.3711    3.9398   0.5143 2.666  0.00768 **
## salarymedium 0.8857    2.4247   0.5231 1.693  0.09041 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## salarylow        3.940     0.2538    1.4377    10.796
## salarymedium     2.425     0.4124    0.8698     6.759
##
## Concordance= 0.584  (se = 0.028 )
## Likelihood ratio test= 14.16  on 2 df,   p=8e-04
## Wald test            = 11.8  on 2 df,   p=0.003
## Score (logrank) test = 12.74  on 2 df,   p=0.002
```

```
##        chisq df    p
## salary 0.747  2 0.69
## GLOBAL 0.747  2 0.69
```

I dipendenti con stipendio basso presentano un rischio di abbandono quasi 4 volte più alto rispetto a chi ha uno stipendio alto (riferimento). Anche chi percepisce uno stipendio medio presenta un rischio più elevato di abbandono rispetto alla fascia di riferimento, ma la differenza non è statisticamente significativa ($Pr(>|z|)>0.05$).

La verifica della proporzionalità dei rischi conferma che il modello di Cox è applicabile a questa variabile: il **p-value pari a 0.69** (>0.05) non consente di rifiutare l'ipotesi nulla H0 che sostiene la proporzionalità dei rischi.

Promozione negli ultimi 5 anni

```
## Call:
## coxph(formula = surv_object ~ hr.df$promotion_last_5years, data = hr.df)
##
##   n= 500, number of events= 120
##
##                                    coef  exp(coef)   se(coef)      z Pr(>|z|)
## hr.df$promotion_last_5years -1.604e+01  1.079e-07  1.783e+03 -0.009    0.993
##
##                             exp(coef) exp(-coef) lower .95 upper .95
## hr.df$promotion_last_5years 1.079e-07    9264213         0       Inf
##
## Concordance= 0.512  (se = 0.004 )
## Likelihood ratio test= 5.75  on 1 df,   p=0.02
## Wald test            = 0  on 1 df,   p=1
## Score (logrank) test = 2.91  on 1 df,   p=0.09
```

```
##                               chisq df p
## hr.df$promotion_last_5years 1.22e-08  1 1
## GLOBAL                      1.22e-08  1 1
```

La variabile *promotion_last_5years* non è significativamente associata al rischio di uscita dall'azienda ($Pr(>|z|)$ alto e IC infinito) suggeriscono che l'effetto di questa variabile è altamente incerto e non significativo.
La proporzionalità dei rischi è verificata, ma la capacità predittiva del modello è quasi casuale (0.512).

Reparto

```
## Call:
## coxph(formula = surv_object ~ hr.df$department)
##
##   n= 500, number of events= 120
##
##                               coef exp(coef) se(coef)      z Pr(>|z|)
## hr.df$departmenthr           1.1452    3.1432   0.6906  1.658   0.0972 .
## hr.df$departmentIT           0.1500    1.1618   0.7304  0.205   0.8373
## hr.df$departmentmanagement  -1.5974    0.2024   1.1555 -1.382   0.1668
## hr.df$departmentmarketing    0.6965    2.0067   0.7075  0.984   0.3249
## hr.df$departmentproduct_mng  0.1446    1.1556   0.7310  0.198   0.8432
## hr.df$departmentRandD       -0.2156    0.8060   0.9135 -0.236   0.8134
## hr.df$departmentsales        0.6088    1.8382   0.6051  1.006   0.3144
## hr.df$departmentsupport      0.9819    2.6695   0.6087  1.613   0.1067
## hr.df$departmenttechnical    0.9825    2.6712   0.6031  1.629   0.1033
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                             exp(coef) exp(-coef) lower .95 upper .95
## hr.df$departmenthr             3.1432     0.3181   0.81197    12.168
## hr.df$departmentIT             1.1618     0.8607   0.27757     4.863
## hr.df$departmentmanagement     0.2024     4.9403   0.02102     1.949
## hr.df$departmentmarketing      2.0067     0.4983   0.50149     8.030
## hr.df$departmentproduct_mng    1.1556     0.8653   0.27582     4.842
## hr.df$departmentRandD          0.8060     1.2406   0.13452     4.830
## hr.df$departmentsales          1.8382     0.5440   0.56149     6.018
## hr.df$departmentsupport        2.6695     0.3746   0.80961     8.802
## hr.df$departmenttechnical      2.6712     0.3744   0.81913     8.711
##
## Concordance= 0.622  (se = 0.03 )
## Likelihood ratio test= 26.29  on 9 df,   p=0.002
## Wald test            = 17.57  on 9 df,   p=0.04
## Score (logrank) test = 21.69  on 9 df,   p=0.01

##                   chisq df    p
## hr.df$department  13.7  9 0.13
## GLOBAL            13.7  9 0.13
```

Nessun reparto aziendale incide in modo molto significativo sul rischio di abbandono, infatti i valori di *Pr(>|z|)* che indicano la probabilità che il valore osservato del test statistico si verifichi sotto H0, sono tutti sopra la soglia di 0.05. I valori più bassi (marginalmente significativi) sono legati ai reparti *HR*, *support* e *technical*, i cui dipendenti presentanto una probabilità di abbandono superiore rispetto a chi lavora nel reparto *account* (riferimento) di circa 2-3 volte.
Anche in questo caso, il test di proporzionalità dei rischi indica che il modello di Cox è adeguato (p-value=0.13).

Infortuni sul lavoro

```
## Call:
## coxph(formula = surv_object ~ hr.df$work_accident, data = hr.df)
##
##   n= 500, number of events= 120
##
```

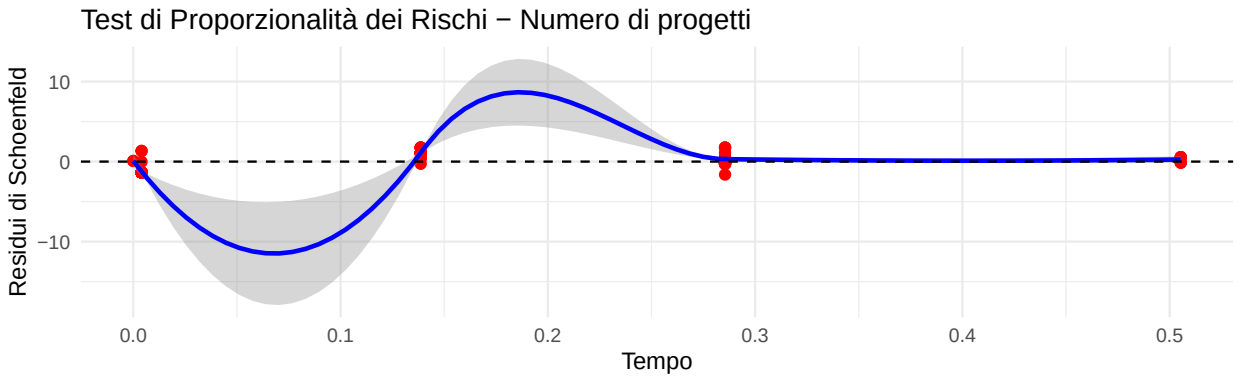
```
##                          coef exp(coef) se(coef)      z Pr(>|z|)
## hr.df$work_accident -1.5469    0.2129   0.5850 -2.644  0.00819 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## hr.df$work_accident    0.2129      4.697   0.06765    0.6701
##
## Concordance= 0.549  (se = 0.011 )
## Likelihood ratio test= 11.9  on 1 df,   p=6e-04
## Wald test            = 6.99  on 1 df,   p=0.008
## Score (logrank) test = 8.5  on 1 df,   p=0.004
```

```
##                       chisq df   p
## hr.df$work_accident 0.0158  1 0.9
## GLOBAL              0.0158  1 0.9
```

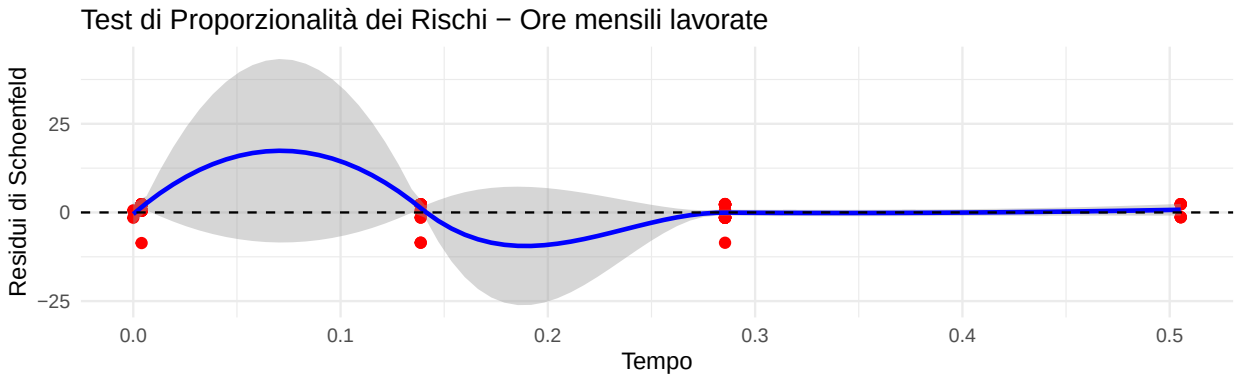
I dipendenti che non hanno avuto un infortunio sul lavoro presentano un rischio di abbandono 5 volte superiore rispetto a chi lo ha subito, il coefficiente risulta significativo (*Pr(>|z|)<0.01*).
La verifica della proporzionalità dei rischi è confermata dal p-value pari a 0.9.

Violazione dell'ipotesi di prorzionalità dei rischi

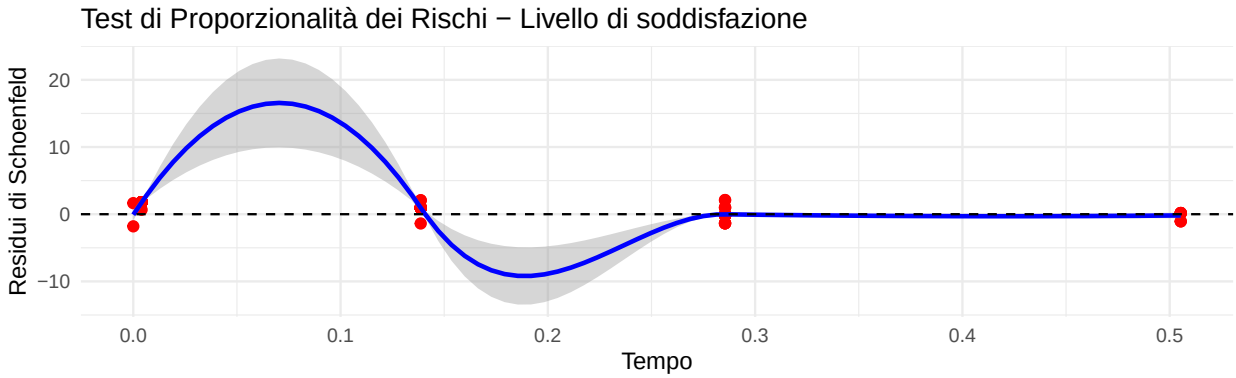
In questa sezione si riportano sinteticamente i risultati della funzione **cox.zph()** di R, utilizzata per eseguire il *Test di Proporzionalità dei Rischi di Schoenfeld* per le covariate per cui l'assunzione di proporzionalità dei rischi è violata (**p-value<0.01**), e per cui il modello di Cox non risulta quindi applicabile.



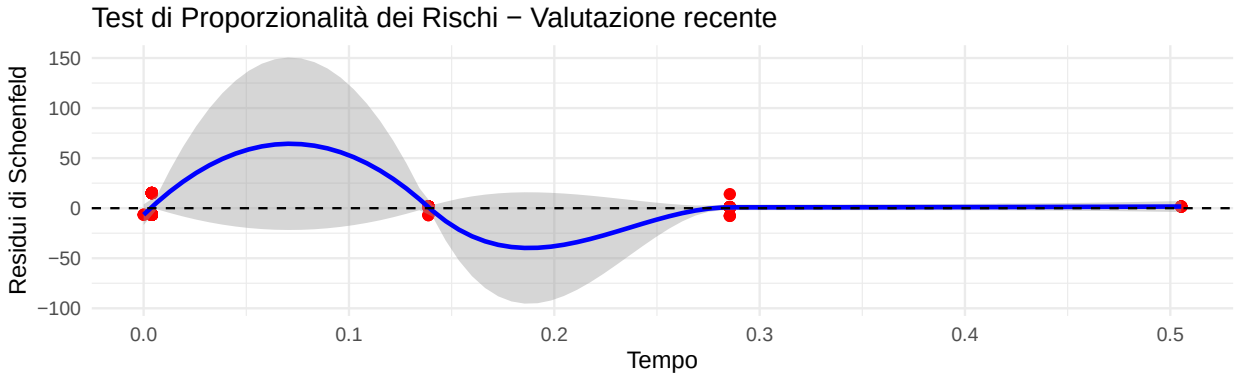

```
##                         chisq df      p
## hr.df$number_project     68.1  1 <2e-16
## GLOBAL                   68.1  1 <2e-16
```




```
##                    chisq df       p
## hr.df$hours_cat     61.6  4 1.3e-12
## GLOBAL              61.6  4 1.3e-12
```




```
##                          chisq df       p
## hr.df$satisfaction_cat  61.5  3 2.8e-13
## GLOBAL                  61.5  3 2.8e-13
```




```
##                         chisq df       p
## hr.df$evaluation_cat  66.2  2 4.1e-15
## GLOBAL                66.2  2 4.1e-15
```

Modello di Cox con più covariate

Si valuta il modello di Cox inserendo tutte le 4 covariate per cui è stata verificata con successo la proporzionalità dei rischi, quindi il livello salariale, il reparto, infortunio e promozione.

```
## Call:
## coxph(formula = surv_object ~ salary + department + work_accident +
##     promotion_last_5years, data = hr.df)
##
##   n= 500, number of events= 120
##
##                             coef  exp(coef)   se(coef)      z Pr(>|z|)
## salarylow              8.372e-01  2.310e+00  5.272e-01  1.588   0.1123
## salarymedium           4.754e-01  1.609e+00  5.319e-01  0.894   0.3714
## departmenthr           1.037e+00  2.821e+00  6.909e-01  1.501   0.1334
## departmentIT          -5.008e-02  9.512e-01  7.315e-01 -0.068   0.9454
## departmentmanagement  -1.192e+00  3.036e-01  1.171e+00 -1.018   0.3085
## departmentmarketing    8.017e-01  2.229e+00  7.083e-01  1.132   0.2577
## departmentproduct_mng -1.389e-01  8.704e-01  7.326e-01 -0.190   0.8497
## departmentRandD       -2.916e-01  7.470e-01  9.148e-01 -0.319   0.7499
## departmentsales        4.418e-01  1.556e+00  6.068e-01  0.728   0.4665
## departmentsupport      7.532e-01  2.124e+00  6.101e-01  1.235   0.2170
## departmenttechnical    8.108e-01  2.250e+00  6.040e-01  1.342   0.1795
## work_accident         -1.330e+00  2.646e-01  5.920e-01 -2.246   0.0247 *
## promotion_last_5years -1.538e+01  2.099e-07  1.839e+03 -0.008   0.9933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## salarylow             2.310e+00  4.329e-01   0.82199    6.4914
## salarymedium          1.609e+00  6.216e-01   0.56721    4.5626
## departmenthr          2.821e+00  3.545e-01   0.72825   10.9242
## departmentIT          9.512e-01  1.051e+00   0.22676    3.9896
## departmentmanagement  3.036e-01  3.294e+00   0.03061    3.0110
## departmentmarketing   2.229e+00  4.486e-01   0.55623    8.9344
## departmentproduct_mng 8.704e-01  1.149e+00   0.20706    3.6585
## departmentRandD       7.470e-01  1.339e+00   0.12435    4.4880
## departmentsales       1.556e+00  6.429e-01   0.47357    5.1093
## departmentsupport     2.124e+00  4.709e-01   0.64235    7.0218
## departmenttechnical   2.250e+00  4.445e-01   0.68867    7.3491
## work_accident         2.646e-01  3.780e+00   0.08292    0.8442
## promotion_last_5years 2.099e-07  4.765e+06   0.00000       Inf
##
## Concordance= 0.674  (se = 0.027 )
## Likelihood ratio test= 45.8  on 13 df,   p=2e-05
## Wald test            = 27.83  on 13 df,   p=0.01
## Score (logrank) test = 36.53  on 13 df,   p=5e-04
```

Il valore di **concordance** di **0.674** indica una moderata capacità del modello di discriminare l'evento. I p-value dei test (Likelihood ratio, Wald, e Score) sono tutti inferiori a 0.05, indicando che il modello è significativo.
Tra i coefficienti, quello risulta più significativo è l'evento di infortunio sul lavoro. L'*odds ratio* pari a 0.2646 conferma che un incidente sul lavoro riduce il rischio di abbandono. Il coefficiente di promozione ha p-value=1 e IC infinito, quindi si considera non significativo nel modello. Il coefficiente per salario basso implica un rischio 2.31 volte più grande rispetto ad un salario alto, anche se il p-value indica una significatività moderata.

Non avendo un impatto significativo, si rivaluta il modello eliminando la variabile legata alla promozione negli ultimi 5 anni. Di seguito il modello di Cox con livello salariale, il reparto e infortunio.

```
## Call:
## coxph(formula = surv_object ~ salary + department + work_accident,
##     data = hr.df)
##
##   n= 500, number of events= 120
##
##                            coef exp(coef)  se(coef)      z Pr(>|z|)
## salarylow              0.962566  2.618406  0.524503  1.835   0.0665 .
## salarymedium           0.586774  1.798179  0.529799  1.108   0.2681
## departmenthr           1.062473  2.893518  0.690891  1.538   0.1241
## departmentIT           0.009034  1.009075  0.731344  0.012   0.9901
## departmentmanagement  -1.240991  0.289097  1.165999 -1.064   0.2872
## departmentmarketing    0.847697  2.334264  0.708112  1.197   0.2313
## departmentproduct_mng -0.084974  0.918536  0.732943 -0.116   0.9077
## departmentRandD       -0.233111  0.792066  0.915211 -0.255   0.7989
## departmentsales        0.482893  1.620757  0.607104  0.795   0.4264
## departmentsupport      0.809920  2.247729  0.610267  1.327   0.1845
## departmenttechnical    0.871900  2.391450  0.603920  1.444   0.1488
## work_accident         -1.340778  0.261642  0.591953 -2.265   0.0235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## salarylow                2.6184     0.3819   0.93665    7.3197
## salarymedium             1.7982     0.5561   0.63660    5.0792
## departmenthr             2.8935     0.3456   0.74703   11.2076
## departmentIT             1.0091     0.9910   0.24066    4.2310
## departmentmanagement     0.2891     3.4590   0.02941    2.8415
## departmentmarketing      2.3343     0.4284   0.58264    9.3518
## departmentproduct_mng    0.9185     1.0887   0.21838    3.8635
## departmentRandD          0.7921     1.2625   0.13174    4.7620
## departmentsales          1.6208     0.6170   0.49312    5.3270
## departmentsupport        2.2477     0.4449   0.67965    7.4337
## departmenttechnical      2.3914     0.4182   0.73216    7.8112
## work_accident            0.2616     3.8220   0.08200    0.8348
##
## Concordance= 0.671  (se = 0.027 )
## Likelihood ratio test= 43.11  on 12 df,   p=2e-05
## Wald test            = 29.6  on 12 df,   p=0.003
## Score (logrank) test = 35.5  on 12 df,   p=4e-04
```

Il modello senza promotion presenta una concordance simile a quella del modello precedente.
Migliora lievemente la significatività della variabile salario (p-value=0.0665).

Poiché il modello di Cox presuppone la proporzionalità dei rischi, non è stato inserito il livello di soddisfazione. Tuttavia, nelle sezioni precedenti è stato classificato come fattore determinante per la permanenza in azienda. Il modello di seguito considera, oltre alle variabili di cui al modello precedente, la *stratificazione* della categoriale del livello di soddisfazione, attraverso la funzione R *strata*. Stratificando alcune variabili, infatti, puoi eliminare la necessità di modellare direttamente il loro effetto sui rischi proporzionali, permettendo al modello di concentrarsi meglio sulle altre covariate.

```
## Call:
## coxph(formula = surv_object ~ salary + department + work_accident +
##     strata(satisfaction_cat), data = hr.df)
##
##   n= 500, number of events= 120
##
##                           coef exp(coef) se(coef)      z Pr(>|z|)
## salarylow              1.3125    3.7153   0.5200  2.524  0.01161 *
## salarymedium           1.0451    2.8436   0.5293  1.975  0.04832 *
## departmenthr           1.3311    3.7852   0.7039  1.891  0.05863 .
## departmentIT          -0.3790    0.6845   0.7372 -0.514  0.60720
## departmentmanagement  -1.1648    0.3120   1.1735 -0.993  0.32093
## departmentmarketing    0.7693    2.1583   0.7209  1.067  0.28593
## departmentproduct_mng -0.2012    0.8178   0.7391 -0.272  0.78550
## departmentRandD       -0.9472    0.3878   0.9348 -1.013  0.31094
## departmentsales        0.2416    1.2733   0.6168  0.392  0.69525
## departmentsupport      0.5801    1.7862   0.6219  0.933  0.35096
## departmenttechnical    0.5935    1.8103   0.6126  0.969  0.33262
## work_accident         -1.7015    0.1824   0.6007 -2.832  0.00462 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## salarylow                3.7153     0.2692   1.34068   10.2958
## salarymedium             2.8436     0.3517   1.00775    8.0241
## departmenthr             3.7852     0.2642   0.95257   15.0412
## departmentIT             0.6845     1.4608   0.16138    2.9037
## departmentmanagement     0.3120     3.2052   0.03128    3.1121
## departmentmarketing      2.1583     0.4633   0.52534    8.8670
## departmentproduct_mng    0.8178     1.2228   0.19208    3.4817
## departmentRandD          0.3878     2.5785   0.06207    2.4230
## departmentsales          1.2733     0.7853   0.38009    4.2659
## departmentsupport        1.7862     0.5598   0.52789    6.0439
## departmenttechnical      1.8103     0.5524   0.54492    6.0142
## work_accident            0.1824     5.4819   0.05620    0.5921
##
## Concordance= 0.746  (se = 0.033 )
## Likelihood ratio test= 46.04  on 12 df,   p=7e-06
## Wald test            = 34.22  on 12 df,   p=6e-04
## Score (logrank) test = 39.19  on 12 df,   p=1e-04
```

La concordance del modello è pari a **0.746**, la più alta tra i modelli finora analizzati.
Presenta inoltre una migliore significatività delle variabili relative a *salary* (low e medium) e *work_accident*, aumentando anche quella del *department* HR. Anche i valori dei test risultano maggiormente significativi rispetto ai modelli precedenti.

Modello di Cox con una nuova variabile (trust)

Dall'analisi svolta finora emerge che la non linearità delle covariate rispetto all'abbandono dell'azienda è principalmente dovuta al carico di lavoro dei dipendenti. Infatti, i dipendenti tendono a lasciare l'azienda non solo quando il carico di lavoro è eccessivo, ma anche quando risulta troppo basso, ovvero quando vengono assegnati pochi progetti e lavorano meno ore.
Questo fenomeno probabilmente si traduce in una minore stima e fiducia nei confronti del dipendente.
Per spiegare meglio tale fenomeno, si è deciso di inserire una nuova variabile nel dataframe, in grado di descrivere il livello di fiducia assegnato al dipendente. In particolare, la variabile **trust** verrà impostata a 0 quando il numero di progetti è inferiore a 3, soglia individuata nell'analisi preliminare oltre la quale il carico di lavoro assume un impatto diretto sull'abbandono.

```
hr.df$trust <- ifelse(hr.df$number_project < 3, 0, 1)
```

Si aggiunge **trust** al modello, eliminando le covariate a rischi non proporzionali.

```
## Call:
## coxph(formula = surv_object ~ salary + department + work_accident +
##     trust, data = hr.df)
##
##   n= 500, number of events= 120
##
##                           coef exp(coef) se(coef)       z Pr(>|z|)
## salarylow              1.54274   4.67741  0.53110   2.905  0.00367 **
## salarymedium           1.26255   3.53443  0.53970   2.339  0.01932 *
## departmenthr           1.83839   6.28639  0.71293   2.579  0.00992 **
## departmentIT          -0.40298   0.66832  0.73485  -0.548  0.58342
## departmentmanagement  -0.29297   0.74604  1.16867  -0.251  0.80205
## departmentmarketing    1.65311   5.22321  0.73006   2.264  0.02355 *
## departmentproduct_mng  0.84538   2.32886  0.75522   1.119  0.26298
## departmentRandD       -0.96108   0.38248  0.92994  -1.033  0.30138
## departmentsales        1.02190   2.77848  0.62250   1.642  0.10067
## departmentsupport      1.56137   4.76534  0.63433   2.461  0.01384 *
## departmenttechnical    1.30118   3.67361  0.61984   2.099  0.03580 *
## work_accident         -0.84652   0.42891  0.59522  -1.422  0.15497
## trust                 -2.58528   0.07537  0.22413 -11.535  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## salarylow               4.67741     0.2138   1.65171    13.246
## salarymedium            3.53443     0.2829   1.22723    10.179
## departmenthr            6.28639     0.1591   1.55438    25.424
## departmentIT            0.66832     1.4963   0.15830     2.822
## departmentmanagement    0.74604     1.3404   0.07551     7.371
## departmentmarketing     5.22321     0.1915   1.24885    21.846
## departmentproduct_mng   2.32886     0.4294   0.53002    10.233
## departmentRandD         0.38248     2.6145   0.06181     2.367
## departmentsales         2.77848     0.3599   0.82022     9.412
## departmentsupport       4.76534     0.2098   1.37453    16.521
## departmenttechnical     3.67361     0.2722   1.09015    12.379
## work_accident           0.42891     2.3315   0.13357     1.377
## trust                   0.07537    13.2670   0.04858     0.117
##
```

```
## Concordance= 0.866  (se = 0.022 )
## Likelihood ratio test= 162.6  on 13 df,   p=<2e-16
## Wald test            = 160.2  on 13 df,   p=<2e-16
## Score (logrank) test = 226.9  on 13 df,   p=<2e-16
```

Nel modello che include **trust**, i risultati evidenziano una notevole capacità discriminante (concordance pari a **0.866**. Questo valore suggerisce che il modello è molto efficace nel distinguere tra i dipendenti che lasciano l'azienda e quelli che vi restano.
Trust risulta la variabile più determinante: il suo coefficiente è pari a -2.58528 con un p-value estremamente basso. L'hazard ratio associato a trust è 0.07537, il che indica che i dipendenti con trust pari a 1 (ossia, quelli a cui viene mostrata fiducia, con almeno 3 progetti assegnati) hanno un rischio di abbandono inferiore di circa il 92,5% rispetto a quelli con trust pari a 0. Questi risultati sottolineano l'importanza della fiducia e, per estensione, della percezione del carico di lavoro e del riconoscimento, come fattori protettivi contro il turnover.
Anche le variabili relative allo stipendio hanno effetti più rilevanti che nei precedenti modelli, ed entrano in gioco per significatività anche le variabili relative ai reparti *HR*, *marketing*, *support* e *technical*.
I test globali del modello (Likelihood Ratio, Wald e Score test) confermano l'elevata significatività complessiva.

Conclusioni

L'analisi di sopravvivenza condotta su questo dataset ha permesso di analizzare i fattori che influenzano la **permanenza dei dipendenti in azienda**. Attraverso l'uso dello stimatore di **Kaplan-Meier**, è stato evidenziato come variabili quali stipendio, promozioni, soddisfazione lavorativa e carico di lavoro abbiano un impatto significativo sulla probabilità di permanenza. I dipendenti con stipendi più bassi tendono a lasciare l'azienda più rapidamente, così come coloro che hanno un carico di lavoro troppo alto o troppo basso.

L'analisi della **regressione di Cox** ha confermato questi risultati, evidenziando ulteriormente il ruolo determinante di alcune covariate. In particolare, l'introduzione della variabile **trust**, definita in funzione del numero di progetti assegnati, ha consentito di catturare l'effetto della fiducia e del riconoscimento sul turnover. Il modello mostra che un valore elevato di trust (ossia quando i dipendenti hanno almeno 3 progetti assegnati) è associato ad una rilevante riduzione del rischio di uscita, suggerendo che sentirsi stimati e valorizzati può essere un importante fattore protettivo contro l'abbandono.

Altri fattori, come il reparto di appartenenza e il verificarsi di infortuni sul lavoro, influenzano il rischio di uscita, sebbene in misura minore. Un aspetto critico dell'analisi è stato il controllo dell'ipotesi di proporzionalità dei rischi, che ha richiesto l'introduzione di stratificazioni per migliorare l'adeguatezza del modello. L'inclusione di covariate multiple, unitamente all'aggiunta della variabile trust, ha permesso di affinare le previsioni e ottenere modelli più robusti e predittivi.

Questo project work rappresenta un'applicazione delle competenze acquisite durante il corso di **Statistical Models**. Inoltre, il percorso di apprendimento è in continua evoluzione nel master: le conoscenze che acquisiremo negli altri moduli offriranno ulteriori spunti e strumenti per perfezionare l'analisi, ampliando le prospettive di ricerca sul tema della permanenza dei dipendenti.