Prpject Work Advanced Statistic Fusco Roberta e Luca Ruocco: WAGEPAN: Salari e Determinanti dell'Occupazione

2025-03-28

WAGEPAN: Salari e Determinanti dell'Occupazione

Questo dataset proviene da uno studio econometrico sui salari e i determinanti dell'occupazione dei giovani uomini. Contiene 4360 osservazioni raccolte tra il 1980 e il 1987, con informazioni personali e lavorative.

Variabili principali

nr: Identificativo della persona (ogni individuo ha un numero univoco)

year: Anno dell'osservazione (1980-1987)

agric: 1 se la persona lavora in agricoltura, 0 altrimenti

black: 1 se la persona è afroamericana, 0 altrimenti

bus: 1 se la persona lavora nel settore business, 0 altrimenti

construc: 1 se la persona lavora nel settore edilizio, 0 altrimenti

ent: 1 se la persona lavora nel settore dell'intrattenimento, 0 altrimenti

exper: Esperienza lavorativa, misurata in anni

fin: 1 se la persona lavora nel settore finanziario, 0 altrimenti

hisp: 1 se la persona è ispanica, 0 altrimenti

poorhlth: 1 se la persona è in cattiva salute, 0 altrimenti

hours: Ore lavorate nell'anno

manuf: 1 se la persona lavora nel settore manifatturiero, 0 altrimenti

married: 1 se la persona è sposata, 0 altrimenti

min: 1 se la persona lavora nel settore minerario, 0 altrimenti

nrthcen: 1 se la persona vive nella regione North Central, 0 altrimenti

nrtheast: 1 se la persona vive nella regione North East, 0 altrimenti

occ1 - occ9 Occupazione della persona, divisa in 9 categorie (Vedi sotto per dettagli)

per: 1 se la persona lavora nel servizio personale, 0 altrimenti

pro: 1 se la persona lavora in una professione tecnica o scientifica, 0 altrimenti

pub: 1 se la persona lavora nella pubblica amministrazione, 0 altrimenti

rur: 1 se la persona vive in una zona rurale, 0 altrimenti

south: 1 se la persona vive nella regione Sud degli Stati Uniti, 0 altrimenti

educ: Anni di istruzione completati

tra: 1 se la persona nel trasporto, 0 altrimenti

trad: 1 se la persona lavora nel commercio, 0 altrimenti

union: 1 se la persona fa parte di un sindacato, 0 altrimenti

lwage: Logaritmo naturale del salario

d81 - d87 Dummy variables: 1 se l'osservazione appartiene a quell'anno specifico (1981-1987), 0 altrimenti

expersq: Quadrato dell'esperienza lavorativa (exper^2)

Le variabili occ1 - occ9 rappresentano diverse categorie di occupazione:

occ1: Dirigenti e amministratori

occ2: Professionisti tecnici (es. ingegneri, scienziati)

occ3: Venditori

occ4: Impiegati d'ufficio

occ5: Operai specializzati

occ6: Operai non specializzati

occ7: Lavoratori dei trasporti

occ8: Agricoltori e lavoratori agricoli

occ9: Lavoratori dei servizi

Obiettivo potenziale del dataset: Studiare l'effetto dell'istruzione, dell'esperienza e dell'appartenenza sindacale sui salari. Può essere usato per analizzare la discriminazione salariale, l'impatto della salute sui guadagni e le differenze regionali nei salari.

```
file_path <- "DataRegression2025_unical.RData"
load(file_path)

data=as.data.frame(wagepan)

descriptive_stats <- summary(data)
print(descriptive_stats)
```

```
##        nr             year          agric             black
##  Min.   :   13   Min.   :1980   Min.   :0.00000   Min.   :0.0000
##  1st Qu.: 2329   1st Qu.:1982   1st Qu.:0.00000   1st Qu.:0.0000
##  Median : 4569   Median :1984   Median :0.00000   Median :0.0000
##  Mean   : 5262   Mean   :1984   Mean   :0.03211   Mean   :0.1156
##  3rd Qu.: 8406   3rd Qu.:1985   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :12548   Max.   :1987   Max.   :1.00000   Max.   :1.0000
##       bus             construc          ent              exper
##  Min.   :0.00000   Min.   :0.000   Min.   :0.00000   Min.   : 0.000
##  1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.: 4.000
##  Median :0.00000   Median :0.000   Median :0.00000   Median : 6.000
##  Mean   :0.07592   Mean   :0.075   Mean   :0.01514   Mean   : 6.515
##  3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:0.00000   3rd Qu.: 9.000
##  Max.   :1.00000   Max.   :1.000   Max.   :1.00000   Max.   :18.000
##       fin               hisp          poorhlth          hours
##  Min.   :0.00000   Min.   :0.000   Min.   :0.00000   Min.   : 120
##  1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:2040
##  Median :0.00000   Median :0.000   Median :0.00000   Median :2080
##  Mean   :0.03693   Mean   :0.156   Mean   :0.01697   Mean   :2191
##  3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:0.00000   3rd Qu.:2414
```

```
##  Max.   :1.00000   Max.   :1.000   Max.   :1.00000   Max.   :4992
##      manuf           married           min            nrthcen
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.2823   Mean   :0.439   Mean   :0.0156   Mean   :0.2578
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##     nrtheast          occ1             occ2              occ3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.1901   Mean   :0.1039   Mean   :0.09151   Mean   :0.05344
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##       occ4             occ5             occ6             occ7
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   :0.1115   Mean   :0.2142   Mean   :0.2021   Mean   :0.09197
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##       occ8              occ9              per               pro
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.01468   Mean   :0.1167   Mean   :0.01674   Mean   :0.07638
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##       pub               rur              south             educ
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   : 3.00
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:11.00
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :12.00
##  Mean   :0.04014   Mean   :0.2039   Mean   :0.3507   Mean   :11.77
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:12.00
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :16.00
##       tra              trad             union            lwage
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :-3.579
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.: 1.351
##  Median :0.0000   Median :0.0000   Median :0.000   Median : 1.671
##  Mean   :0.0656   Mean   :0.2681   Mean   :0.244   Mean   : 1.649
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.: 1.991
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   : 4.052
##       d81              d82              d83              d84
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.000   Median :0.000   Median :0.000   Median :0.000
##  Mean   :0.125   Mean   :0.125   Mean   :0.125   Mean   :0.125
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000
##  Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
##       d85              d86              d87            expersq
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :  0.00
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.: 16.00
##  Median :0.000   Median :0.000   Median :0.000   Median : 36.00
```

```
##  Mean   :0.125   Mean   :0.125   Mean   :0.125   Mean   : 50.42
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.: 81.00
##  Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :324.00
```

```r
library(dplyr)

# Creazione della variabile categoriale per l'etnia basata su black e hisp
data$ethnicity <- apply(data[, c("black", "hisp")], 1, function(x) {
  if (sum(x, na.rm = TRUE) == 1) {
    if (x["black"] == 1) {
      return("Black")
    } else if (x["hisp"] == 1) {
      return("Hispanic")
    }
  } else if (sum(x, na.rm = TRUE) == 0) {
    return("White/Other")
  } else {
    return(NA)
  }
})

# Convertire la variabile in fattore con etichette descrittive
data$ethnicity <- factor(data$ethnicity, levels = c("White/Other", "Black",
                                                    "Hispanic"))


# Creazione della variabile categoriale basata su occ1 - occ9
data$occupation <- apply(data[, grep("occ[1-9]", colnames(data))], 1,
                         function(x)
  {
  ifelse(sum(x) == 1, which(x == 1), NA)
})

# Convertire la variabile in fattore con etichette descrittive
data$occupation <- factor(data$occupation, levels = 1:9,
                          labels = c("Dirigenti e amministratori",
                                     "Professionisti tecnici",
                                     "Venditori",
                                     "Impiegati d'ufficio",
                                     "Operai specializzati",
                                     "Operai non specializzati",
                                     "Lavoratori dei trasporti",
                                     "Agricoltori e lavoratori agricoli",
                                     "Lavoratori dei servizi"))


# Selezioniamo solo le colonne di interesse
wagepan_mer <- data%>%
  select(agric, bus, construc, ent, fin, manuf, min, per, pro, pub, trad, tra)%>%
  mutate(sum_binary = rowSums(.,))
# Aggiungiamo una colonna di somma delle variabili

# Contiamo le osservazioni con sum_binary > 1
num_obs <- sum(wagepan_mer$sum_binary > 1)
```

```r
# Creazione della variabile categoriale basata su agric, bus,
# construct, ent, fin, manuf, min, per, pro
data$sector <- apply(data[, c("agric", "bus", "construc", "ent", "fin",
                              "manuf", "min", "per", "pro", "pub","tra","trad")],
                     1, function(x) {
  ifelse(sum(x) == 1, which(x == 1), NA)
})

# Convertire la variabile in fattore con etichette descrittive
data$sector <- factor(data$sector, levels = 1:12,
                      labels = c("Agricoltura", "Business", "Edilizia",
                                 "Intrattenimento", "Finanza",
                                 "Manifatturiero", "Minerario", "Servizio_Personale",
                                 "Professioni tecniche",
                                 "Pubblica_Amministrazione", "Trasporto",
                                 "Commercio"))

# Creazione della somma delle variabili occ1 - occ9 per ogni riga al fine
# di verificare che queste variabili per riga possono al massimo avere un valore 1
occ_sum <- as.data.frame(rowSums(data[, grep("occ[1-9]",
                                             colnames(data))], na.rm = TRUE))
# Stampare il valore massimo di occ_sum
max_occ_sum <- max(occ_sum, na.rm = TRUE)
print(paste("Il valore massimo di occ_sum è:", max_occ_sum))
```

```
## [1] "Il valore massimo di occ_sum è: 1"
```

```r
# Creazione della somma delle variabili agric, bus, construct, ent, fin,
# manuf, min, per, pro per ogni riga al fine di verificare che queste variabili
# per riga possono al massimo avere un valore 1
sect_sum <- rowSums(data[, c("agric", "bus", "construc", "ent", "fin", "manuf",
                             "min", "per", "pro", "pub", "tra","trad")],
                    na.rm = TRUE)
# Stampare il valore massimo di sect_sum
max_industry_sum <- max(sect_sum, na.rm = TRUE)
print(paste("Il valore massimo di industry_sum è:", max_industry_sum))
```

```
## [1] "Il valore massimo di industry_sum è: 1"
```

```r
# Controllare i valori delle nuove variabili
print(table(data$occupation))
```

```
##
##        Dirigenti e amministratori            Professionisti tecnici
##                               453                               399
##                         Venditori               Impiegati d'ufficio
##                               233                               486
##              Operai specializzati          Operai non specializzati
##                               934                               881
##          Lavoratori dei trasporti Agricoltori e lavoratori agricoli
##                               401                                64
##            Lavoratori dei servizi
##                               509
```

```r
print(table(data$sector))
```

```
##
##              Agricoltura                 Business                Edilizia
##                      140                      331                     327
##          Intrattenimento                  Finanza          Manifatturiero
##                       66                      161                    1231
##                Minerario       Servizio_Personale     Professioni tecniche
##                       68                       73                     333
## Pubblica_Amministrazione                Trasporto               Commercio
##                      175                      286                    1169
```

```r
# Assicurarsi che nrthcen, nrtheast e south siano numerici
data$nrthcen <- as.numeric(data$nrthcen)
data$nrtheast <- as.numeric(data$nrtheast)
data$south <- as.numeric(data$south)

# Propagare il valore di nrthcen, nrtheast e south per ogni individuo
data <- data %>%
  group_by(nr) %>%
  mutate(
    nrthcen = ifelse(any(nrthcen == 1, na.rm = TRUE), 1, 0),
    nrtheast = ifelse(any(nrtheast == 1, na.rm = TRUE), 1, 0),
    south = ifelse(any(south == 1, na.rm = TRUE), 1, 0)
  ) %>%
  ungroup()

# Creare la variabile categoriale region
data$region <- ifelse(data$nrthcen == 1, "North Central",
                      ifelse(data$nrtheast == 1, "North East",
                             ifelse(data$south == 1, "South", "Altro")))

# Convertire in fattore
data$region <- factor(data$region, levels = c("North Central", "North East",
                                              "South", "Altro"))


# Controlliamo la distribuzione di lwage
summary(data$lwage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3.579   1.351   1.671   1.649   1.991   4.052
```

```r
# Controllare la distribuzione di hours
summary(data$hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     120    2040    2080    2191    2414    4992
```

```r
# Stimare il reddito annuo considerando che lwage è il logaritmo
# naturale del salario orario (exp(data$lwage))
data$income_estimated = exp(data$lwage) * data$hours

# Controllare la distribuzione dell'income stimato
summary(data$income_estimated)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    55.24  8270.68 11888.11 12910.40 16541.35 69930.77
```

```r
# Creazione del dataset ridotto eliminando le variabili occ1 - occ9 e agric, bus,
# construc, ent, fin, manuf, min, per, pro, nrthcen, nrtheast, south, black, hisp,
# le variabili dummy d81 - d87 e expersq
data_rid <- data %>% select(-c(occ1:occ9, agric, bus, construc, ent, fin,
                               manuf, min, per, pro, nrthcen, nrtheast, south,
                               black, hisp, pub, tra, trad, d81:d87,expersq))

# Stampare le prime righe del nuovo dataset ridotto
print(head(data_rid))
```

```
## # A tibble: 6 x 15
##      nr  year exper poorhlth hours married   rur  educ union lwage ethnicity
##   <int> <int> <int>    <int> <int>   <int> <int> <int> <int> <dbl> <fct>
## 1    13  1980     1        0  2672       0     0    14     0  1.20 White/Other
## 2    13  1981     2        0  2320       0     0    14     1  1.85 White/Other
## 3    13  1982     3        0  2940       0     0    14     0  1.34 White/Other
## 4    13  1983     4        0  2960       0     0    14     0  1.43 White/Other
## 5    13  1984     5        0  3071       0     0    14     0  1.57 White/Other
## 6    13  1985     6        0  2864       0     0    14     0  1.70 White/Other
## # i 4 more variables: occupation <fct>, sector <fct>, region <fct>,
## #   income_estimated <dbl>
```

```r
# Salvare il dataset ridotto in un file Excel
library(writexl)
write_xlsx(data_rid, "data_rid.xlsx")
```

Il dataset ristrutturato contiene le seguenti variabili:

Variabili Principali

Variabile	Descrizione	Possibili Valori
nr	Identificativo univoco dell'individuo	Numero intero
year	Anno di osservazione	1980 - 1987
exper	Anni di esperienza lavorativa	Numero intero
poorhlth	Indica se l'individuo è in cattiva salute	1 = Sì, 0 = No
hours	Ore lavorate nell'anno	Numero intero
married	Indica se l'individuo è sposato	1 = Sì, 0 = No
rur	Indica se l'individuo vive in una zona rurale	1 = Sì, 0 = No
educ	Anni di istruzione completati	Numero intero
union	Indica se l'individuo è iscritto a un sindacato	1 = Sì, 0 = No
lwage	Logaritmo naturale del salario	Numero decimale
income_estimated	stipendio stimato	Numero decimale

Variabili Categoriali Aggiunte

Le seguenti variabili categoriali sono state derivate per migliorare l'analisi.

Occupazione (`occupation`)

Questa variabile categoriale è stata creata a partire dalle variabili `occ1` - `occ9`.
Possibili valori: - **Dirigenti e amministratori** → Manager e amministratori aziendali
- **Professionisti tecnici** → Tecnici e professionisti altamente qualificati
- **Venditori** → Addetti alle vendite
- **Impiegati d'ufficio** → Lavoratori in ufficio e amministrazione
- **Operai specializzati** → Lavoratori con competenze tecniche specializzate
- **Operai non specializzati** → Lavoratori senza formazione tecnica avanzata
- **Lavoratori dei trasporti** → Persone impiegate nel settore dei trasporti
- **Agricoltori e lavoratori agricoli** → Lavoratori nel settore agricolo
- **Lavoratori dei servizi** → Personale impiegato in servizi generali

---

Settore Economico (`sector`)

Questa variabile categoriale è stata creata utilizzando le variabili `bus`, `construc`, `ent`, `fin`, `manuf`, `min`, `per`, `pro`.
Possibili valori: - **Business** → Settore imprenditoriale e commerciale
- **Edilizia** → Costruzioni e lavori pubblici
- **Intrattenimento** → Cinema, televisione, spettacoli dal vivo
- **Finanza** → Banche, assicurazioni e investimenti
- **Manifatturiero** → Industria manifatturiera e produzione
- **Minerario** → Estrazione mineraria e risorse naturali
- **Servizio_Personale** → Servizio personale - **Professioni tecniche** → Lavoratori in settori scientifici e tecnologici
- **Pubblica Amministrazione** → Lavoratori della Pubblica amministrazione - **Trasporto** → Lavoratori del Trasporto - **Commercio** → Lavoratori del Commercio —

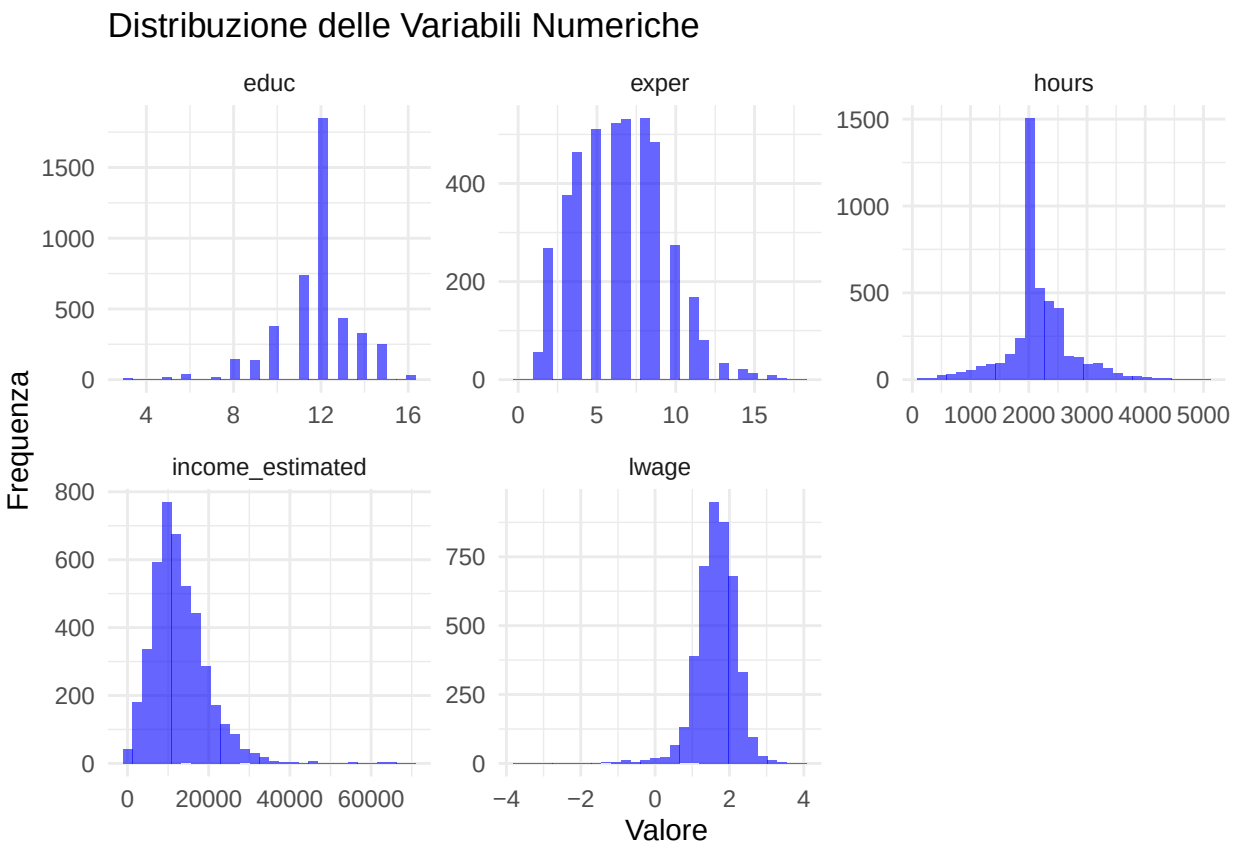
Regione (`region`)

Questa variabile categoriale è stata creata utilizzando `nrthcen`, `nrtheast`, `south`.
Possibili valori: - **North Central** → L'individuo vive nella regione centro-settentrionale degli USA (`nrthcen == 1`)
- **North East** → L'individuo vive nella regione nord-orientale degli USA (`nrtheast == 1`)
- **South** → L'individuo vive nella regione meridionale degli USA (`south == 1`)
- **Altro** → L'individuo vive in una regione non classificata sopra

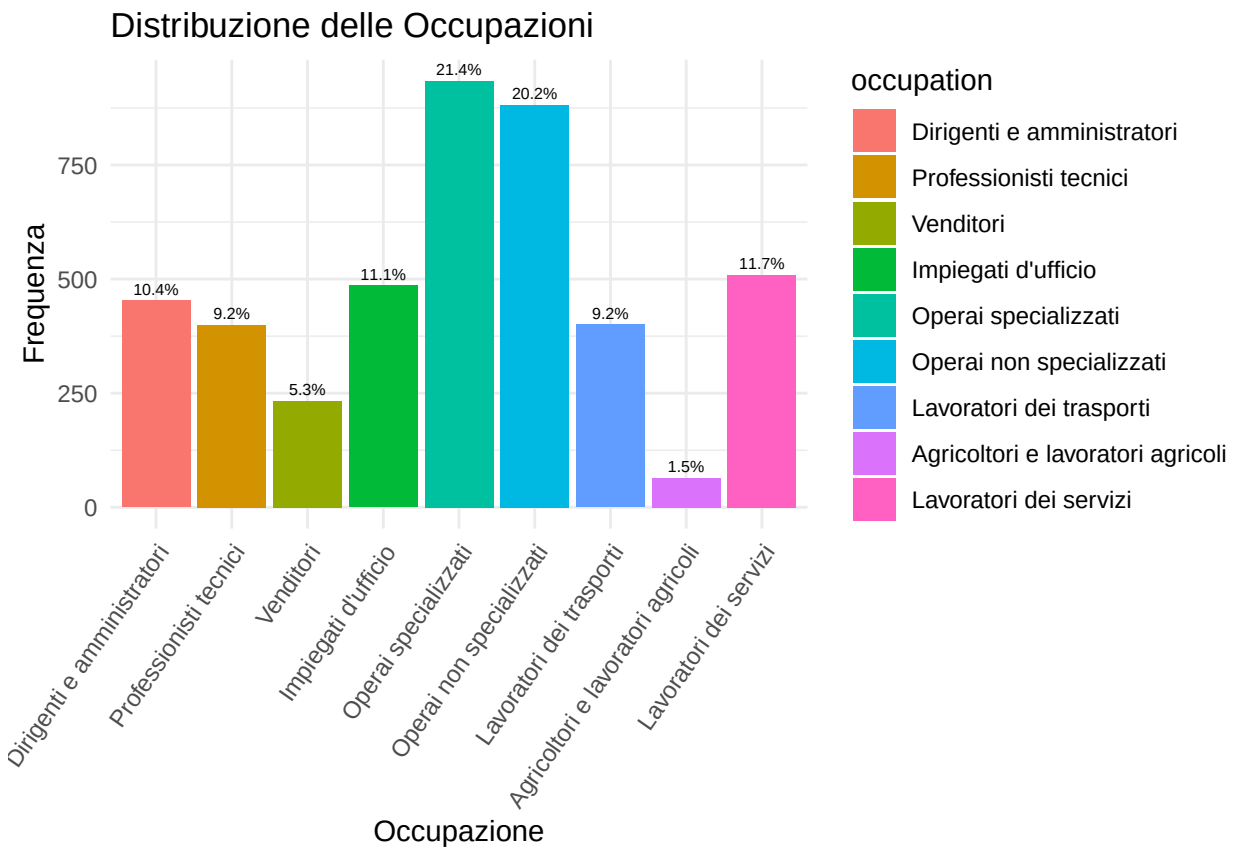
---

```r
library(ggplot2)
library(ggpubr)
library(corrplot)
library(ggcorrplot)
library(tidyr)
library(dplyr)

# Istogrammi per le variabili numeriche
numeric_vars <- c("exper", "hours", "educ", "lwage", "income_estimated")

# Convertire il dataset in formato long per ggplot
data_long <- data_rid %>%
  select(all_of(numeric_vars)) %>%
  pivot_longer(cols = everything(), names_to = "Variabile",
               values_to = "Valore") %>%
```

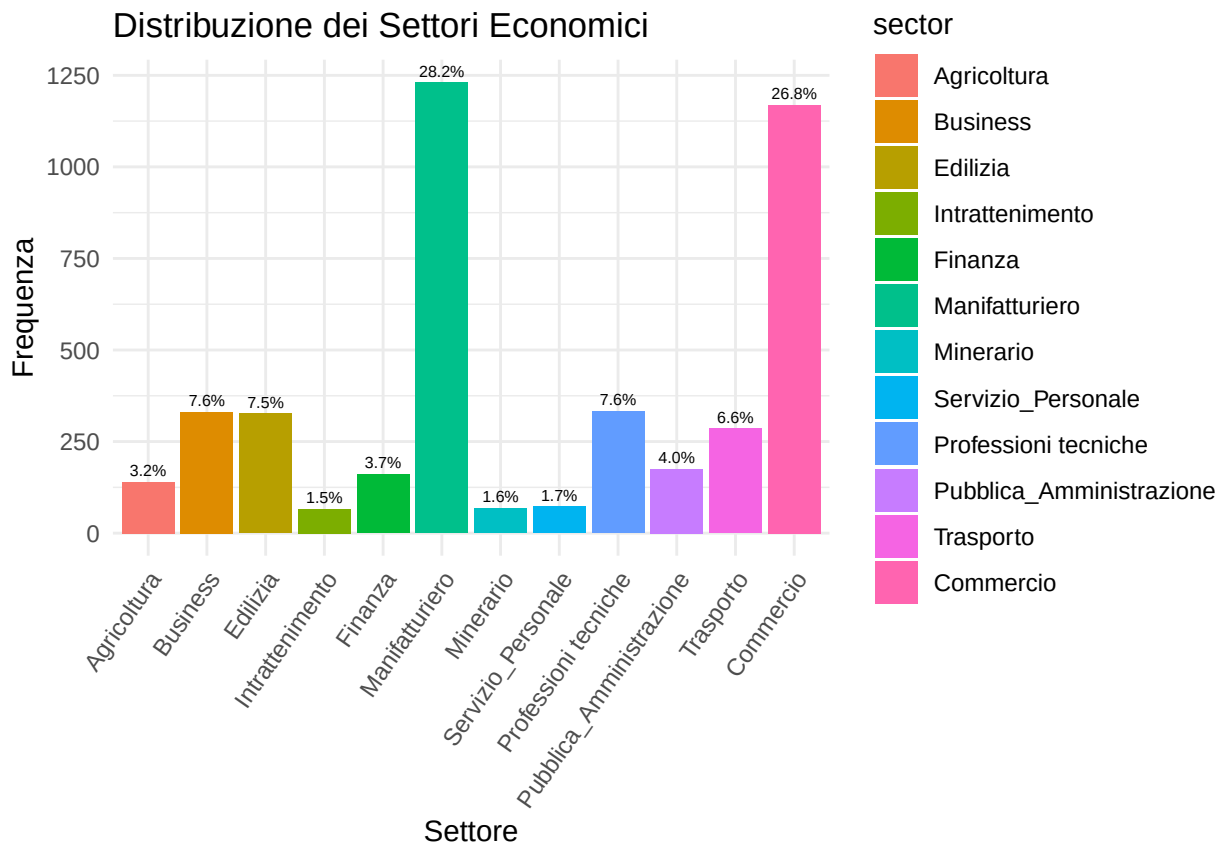
```r
  drop_na()  # Rimuove eventuali NA

# Creare l'istogramma
ggplot(data_long, aes(x = Valore)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.6) +
  facet_wrap(~Variabile, scales = "free") +
  theme_minimal() +
  labs(title = "Distribuzione delle Variabili Numeriche",
       x = "Valore", y = "Frequenza")
```



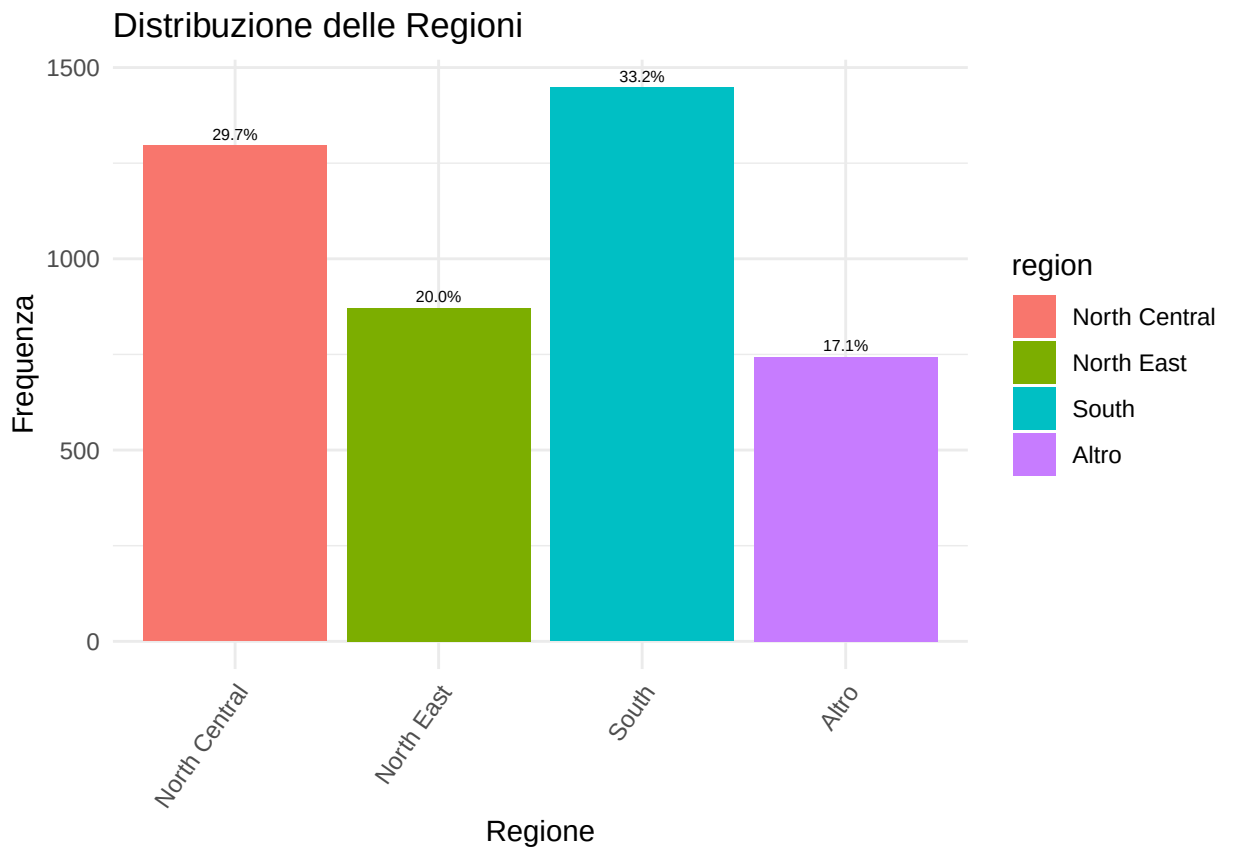

```r
# Analisi delle Variabili Categoriali
# Creazione di un tema personalizzato per i grafici
custom_theme <- theme_minimal() +
  theme(axis.text.x = element_text(angle = 55, hjust = 1))

# Analisi della Distribuzione delle Occupazioni
data_rid %>%
  count(occupation) %>%
  mutate(percentage = n / sum(n) * 100) %>%
  ggplot(aes(x = occupation, y = n, fill = occupation)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = sprintf("%.1f%%", percentage)), vjust = -0.5, size = 2) +
  custom_theme +
  labs(title = "Distribuzione delle Occupazioni", x = "Occupazione",
       y = "Frequenza")
```



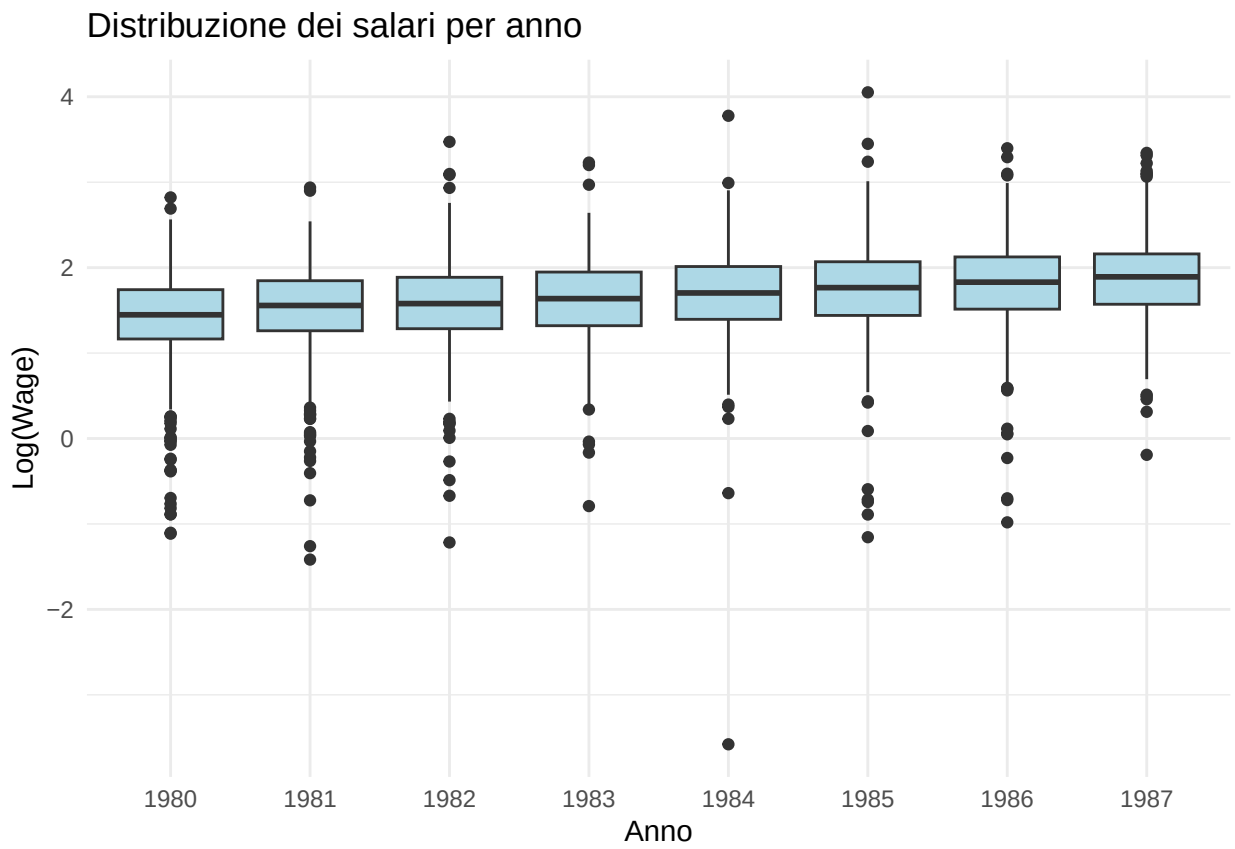

```r
# Analisi della Distribuzione dei Settori Economici
data_rid %>%
  count(sector) %>%
  mutate(percentage = n / sum(n) * 100) %>%
  ggplot(aes(x = sector, y = n, fill = sector)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = sprintf("%.1f%%", percentage)), vjust = -0.5, size = 2) +
  custom_theme +
  labs(title = "Distribuzione dei Settori Economici", x = "Settore",
       y = "Frequenza")
```



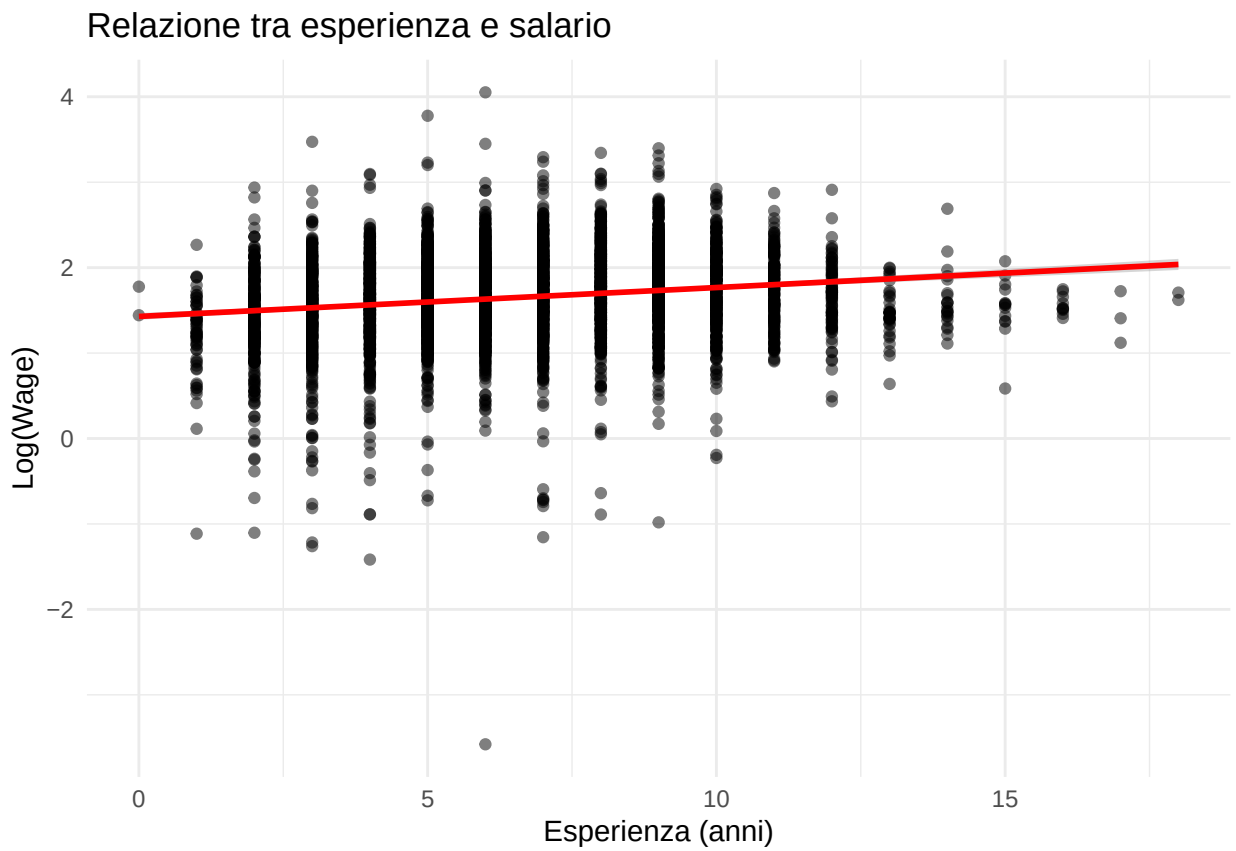

```r
# Analisi della Distribuzione delle Regioni
data_rid %>%
  count(region) %>%
  mutate(percentage = n / sum(n) * 100) %>%
  ggplot(aes(x = region, y = n, fill = region)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = sprintf("%.1f%%", percentage)), vjust = -0.5, size = 2) +
  custom_theme +
  labs(title = "Distribuzione delle Regioni", x = "Regione",
       y = "Frequenza")
```



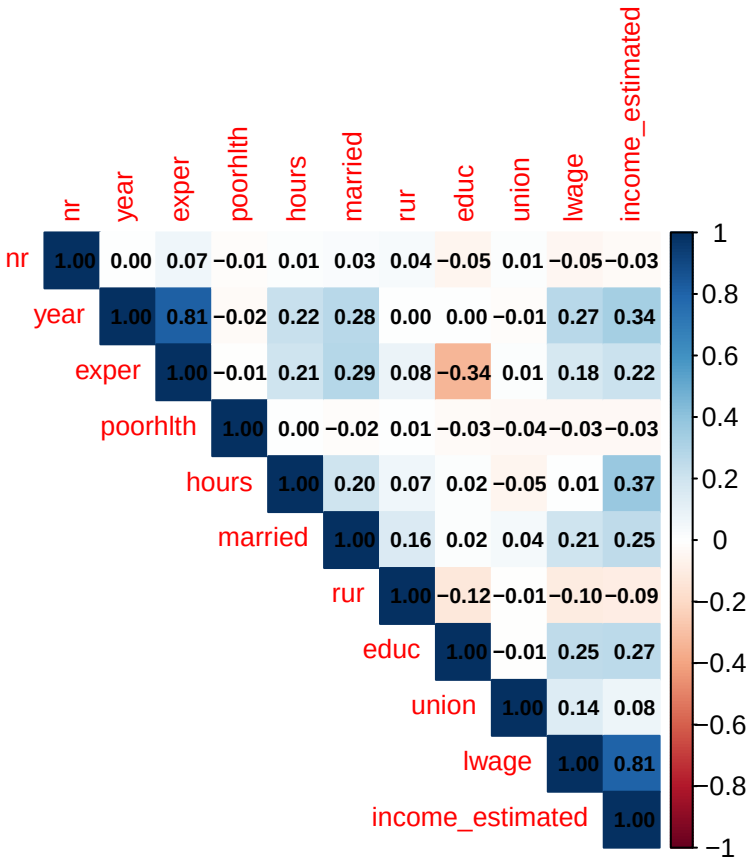

```r
# Boxplot dei salari per anno
ggplot(data_rid, aes(x = as.factor(year), y = lwage)) +
  geom_boxplot(fill = "lightblue") +
  theme_minimal() +
  labs(title = "Distribuzione dei salari per anno", x = "Anno", y = "Log(Wage)")
```



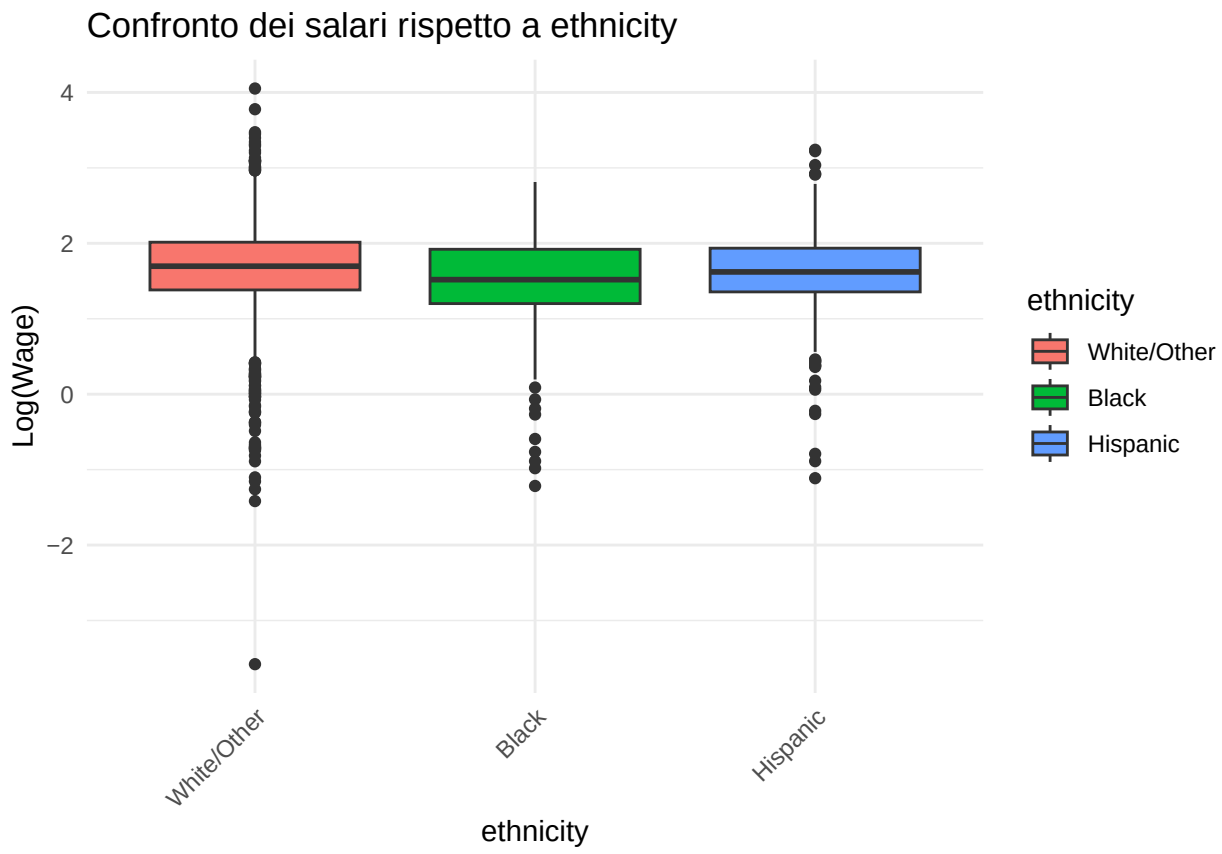

```r
# Relazione tra esperienza e salario
ggplot(data_rid, aes(x = exper, y = lwage)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", col = "red") +
  theme_minimal() +
  labs(title = "Relazione tra esperienza e salario",
       x = "Esperienza (anni)", y = "Log(Wage)")
```



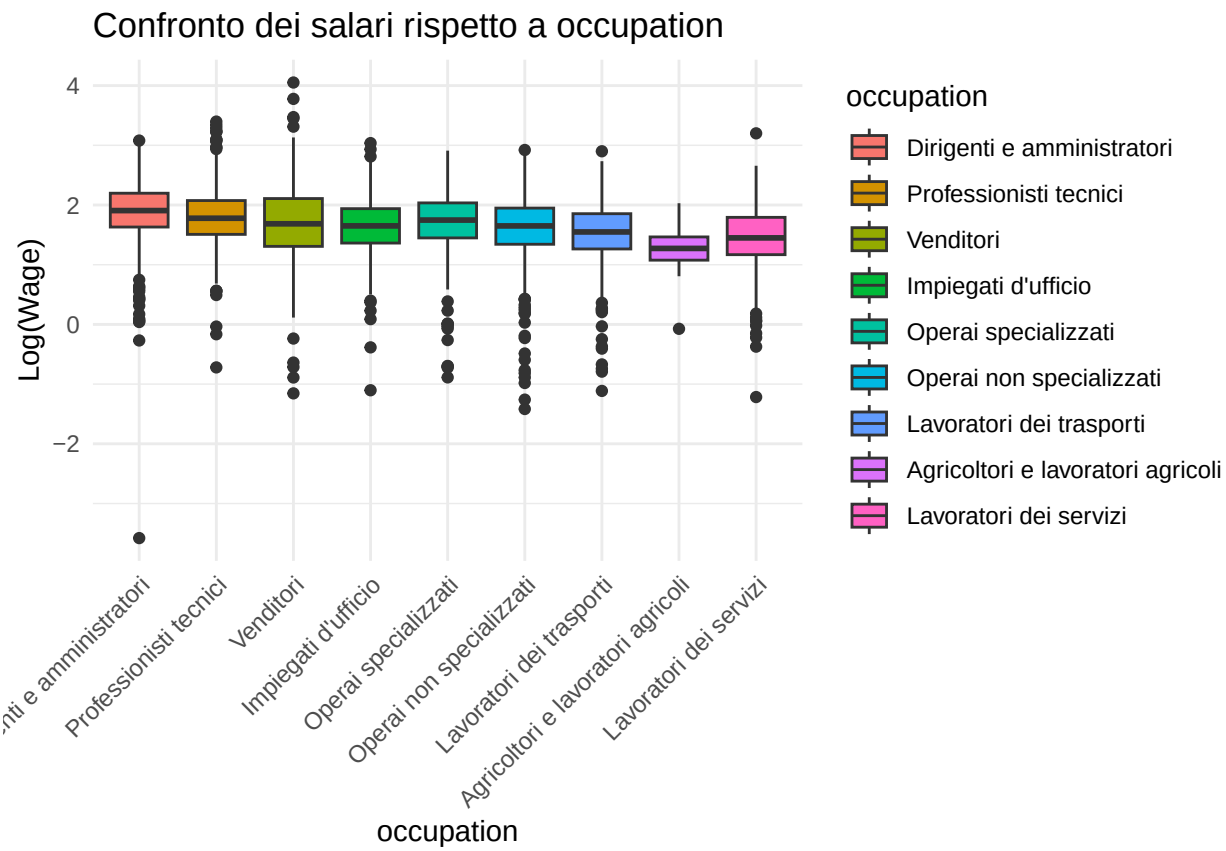

```r
# Matrice di correlazione delle variabili numeriche
# Selezionare solo le variabili numeriche dal dataset
numeric_vars <- data_rid[, sapply(data_rid, is.numeric)]

# Calcolare la matrice di correlazione
corr_matrix <- cor(numeric_vars, use = "complete.obs")

# Creare la matrice di correlazione
corrplot(corr_matrix, method = "color", type = "upper",
         tl.cex = 0.8, addCoef.col = "black", number.cex = 0.7)
```



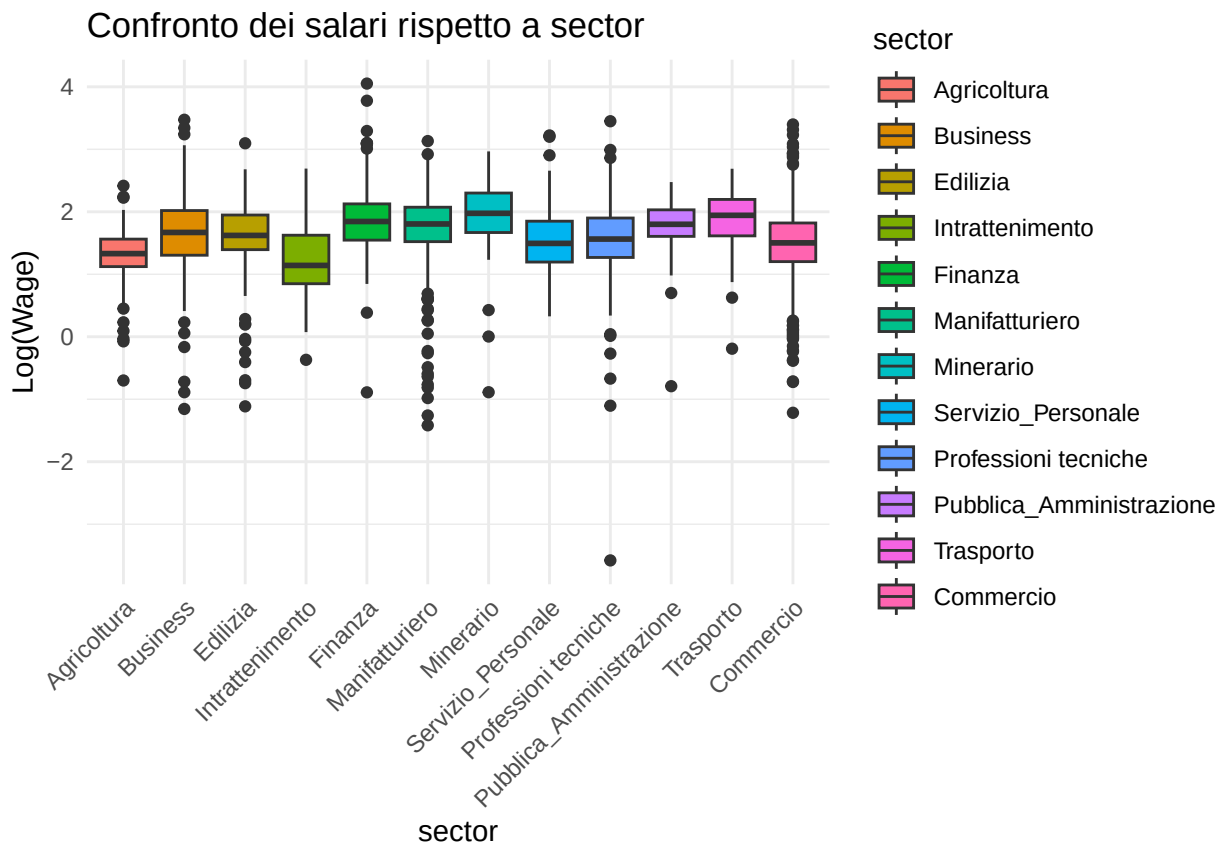

```r
# Lista delle variabili categoriali da analizzare
categorical_vars <- c("ethnicity", "occupation", "sector")

# Creazione di grafici boxplot e test di Kruskal-Wallis
for (var in categorical_vars) {

  # Boxplot per ciascuna variabile categoriale
  p <- ggplot(data_rid, aes_string(x = var, y = "lwage", fill = var)) +
    geom_boxplot() +
    theme_minimal() +
    labs(title = paste("Confronto dei salari rispetto a", var),
         x = var, y = "Log(Wage)") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  print(p)

  # Test di Kruskal-Wallis
  kruskal_test <- kruskal.test(as.formula(paste("lwage ~", var)), data = data_rid)
  print(kruskal_test)
}
```




```
##
##  Kruskal-Wallis rank sum test
##
## data:  lwage by ethnicity
## Kruskal-Wallis chi-squared = 44.413, df = 2, p-value = 2.269e-10
```

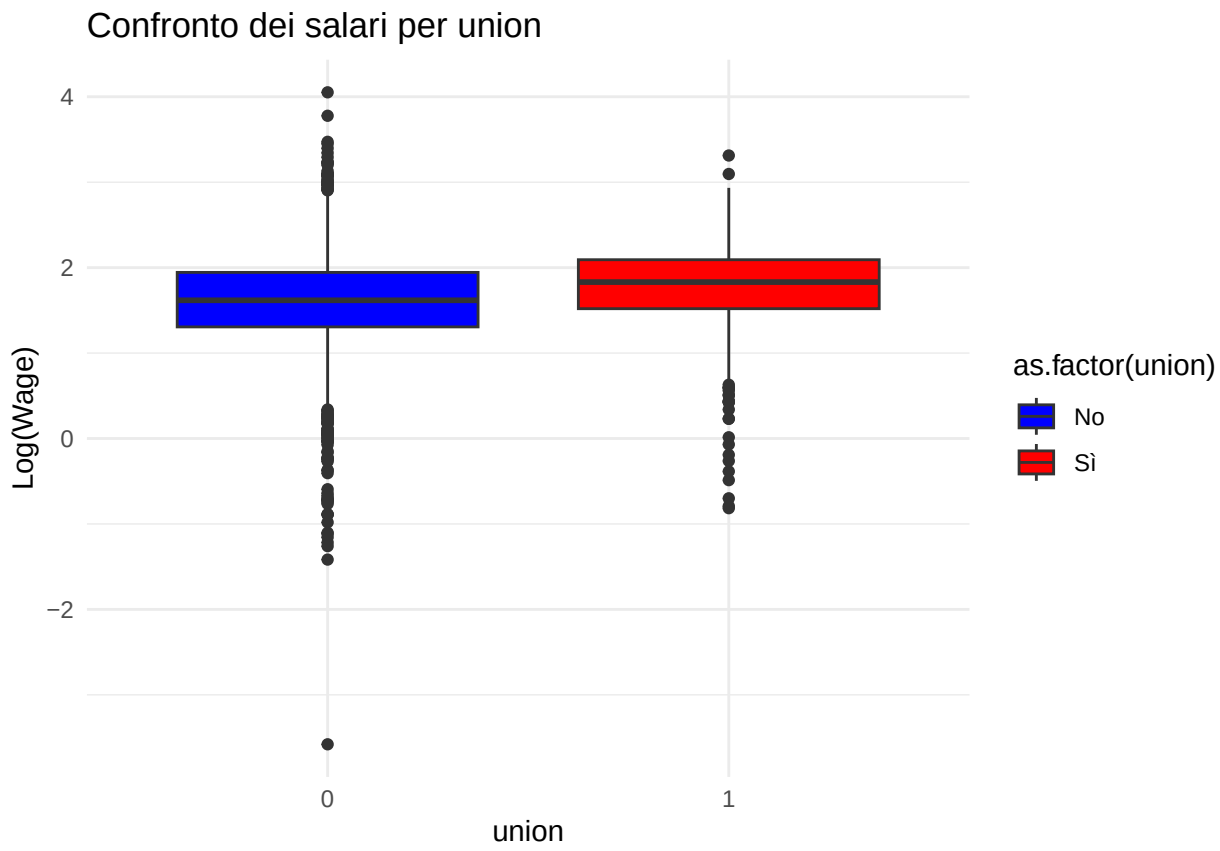



```
##
##  Kruskal-Wallis rank sum test
##
## data:  lwage by occupation
## Kruskal-Wallis chi-squared = 315.09, df = 8, p-value < 2.2e-16
```

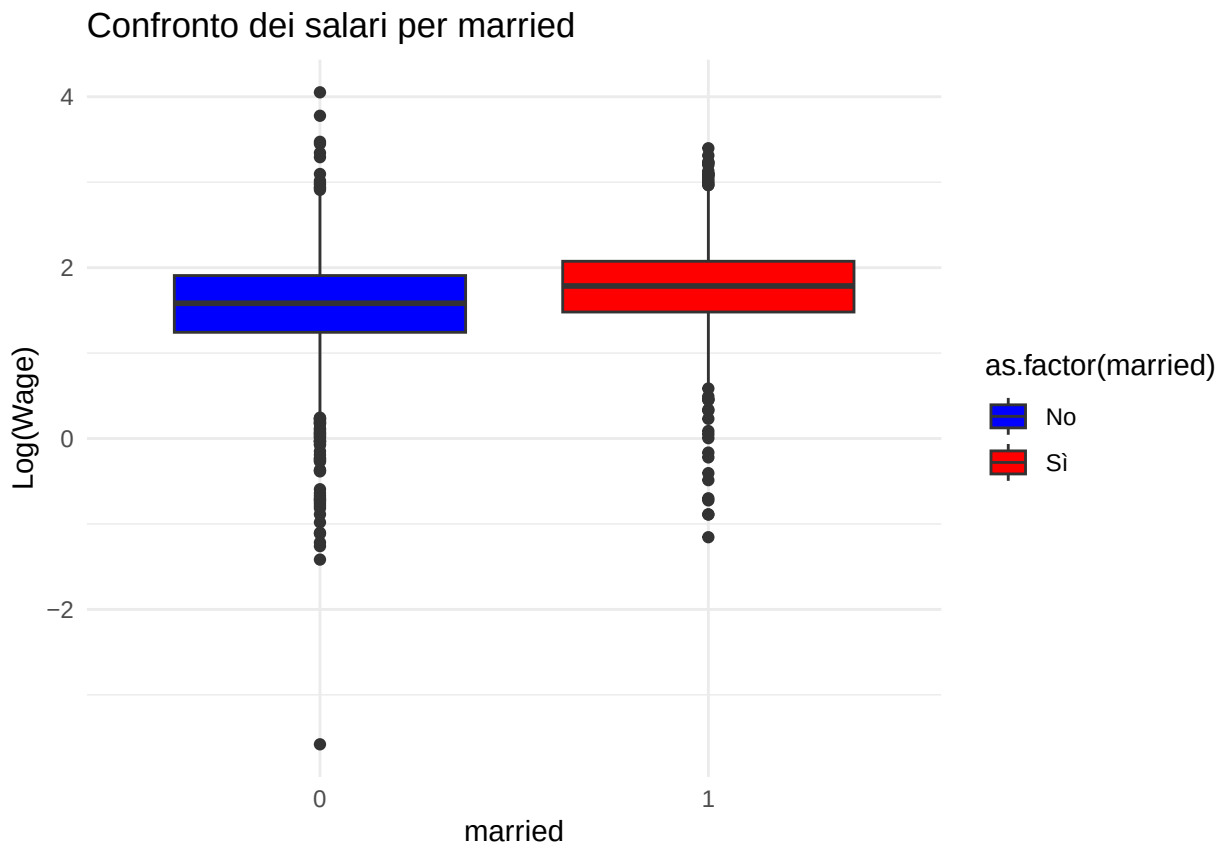



```
##
##  Kruskal-Wallis rank sum test
##
## data:  lwage by sector
## Kruskal-Wallis chi-squared = 509.47, df = 11, p-value < 2.2e-16
```

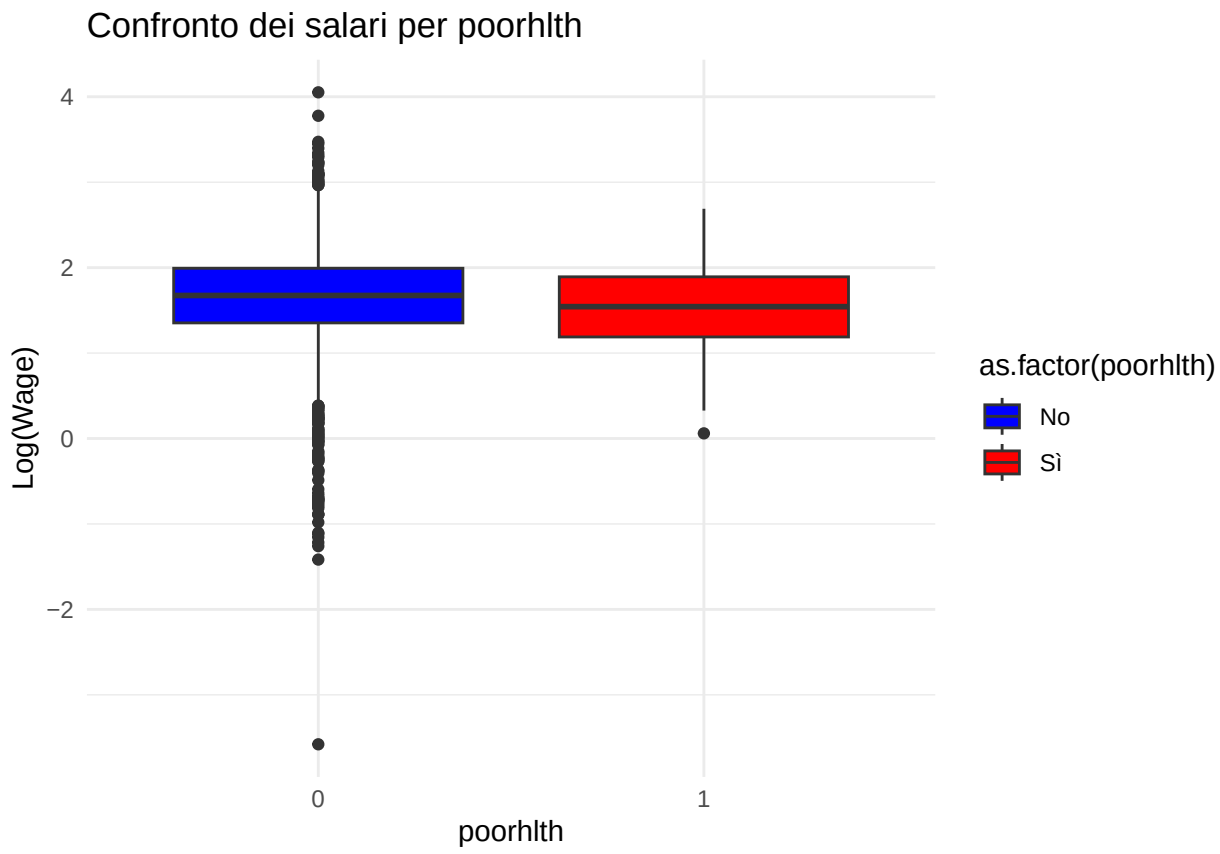
```r
# Lista delle variabili categoriali binarie
binary_vars <- c("union", "married", "poorhlth", "rur")

# Creazione dei boxplot per ciascuna variabile binaria
for (var in binary_vars) {
  p <- ggplot(data_rid, aes(x = as.factor(!!sym(var)), y = lwage,
                            fill = as.factor(!!sym(var)))) +
    geom_boxplot() +
    theme_minimal() +
    labs(title = paste("Confronto dei salari per", var),
         x = var,
         y = "Log(Wage)") +
    scale_fill_manual(values = c("0" = "blue", "1" = "red"), labels = c("No", "Sì"))

  print(p)

  # Test di Kruskal-Wallis
  kruskal_test <- kruskal.test(lwage ~ as.factor(get(var)), data = data_rid)
  print(paste("Test di Kruskal-Wallis per", var))
  print(kruskal_test)
}
```



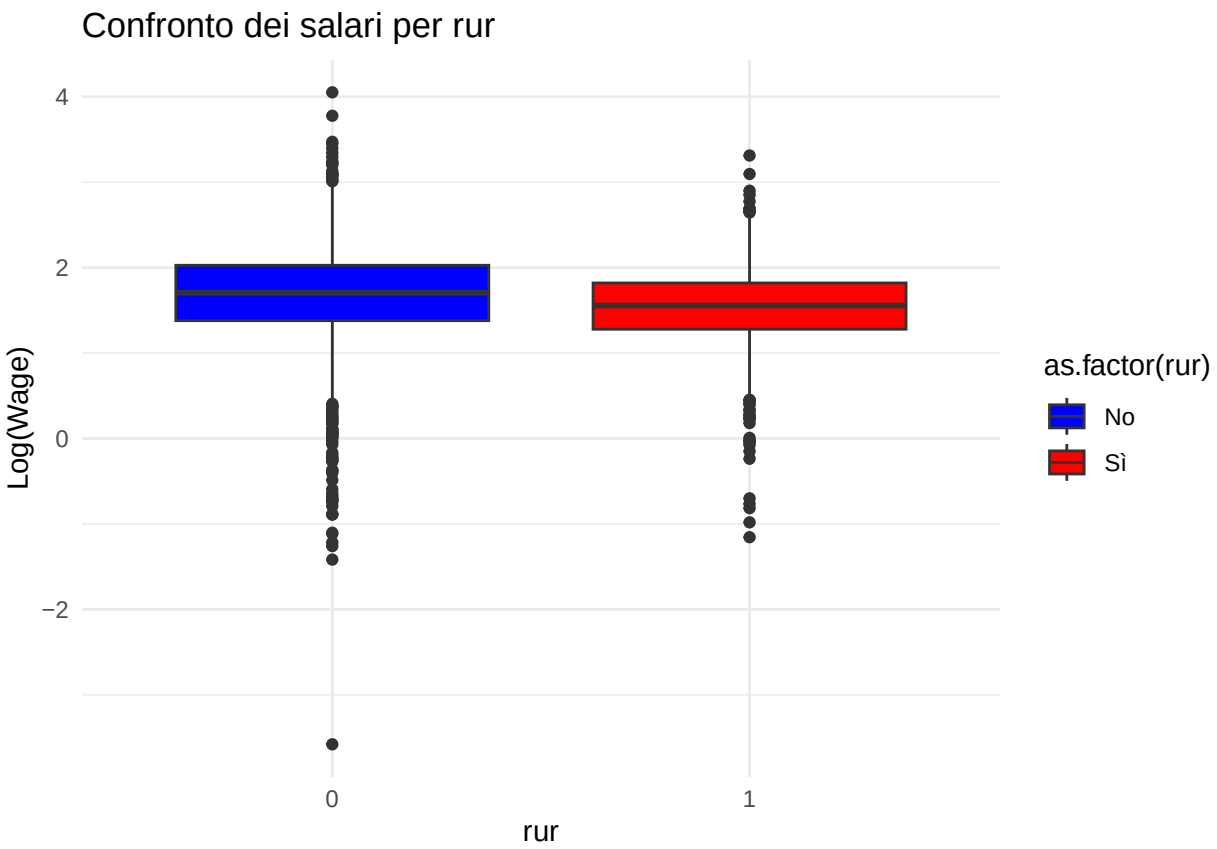

```
## [1] "Test di Kruskal-Wallis per union"
##
##  Kruskal-Wallis rank sum test
##
## data:  lwage by as.factor(get(var))
## Kruskal-Wallis chi-squared = 127.73, df = 1, p-value < 2.2e-16
```




```
## [1] "Test di Kruskal-Wallis per married"
##
##  Kruskal-Wallis rank sum test
##
## data:  lwage by as.factor(get(var))
## Kruskal-Wallis chi-squared = 193.61, df = 1, p-value < 2.2e-16
```




```
## [1] "Test di Kruskal-Wallis per poorhlth"
##
##  Kruskal-Wallis rank sum test
##
## data:  lwage by as.factor(get(var))
## Kruskal-Wallis chi-squared = 4.5989, df = 1, p-value = 0.03199
```




```
## [1] "Test di Kruskal-Wallis per rur"
##
##  Kruskal-Wallis rank sum test
##
## data:  lwage by as.factor(get(var))
## Kruskal-Wallis chi-squared = 65.724, df = 1, p-value = 5.188e-16
```

La distribuzione di educ mostra un picco evidente a 12 anni, suggerendo che molti individui hanno completato la scuola secondaria. Valori più bassi e più alti sono meno frequenti, indicando che meno persone hanno un'istruzione inferiore o superiore alla media.

La distribuzione di exp è asimmetrica a sinistra, con la maggior parte degli individui aventi tra 5 e 15 anni di esperienza lavorativa. Ci sono poche osservazioni per chi ha esperienza prossima a 0 o oltre i 15 anni.

La distribuzione di hours è centrata intorno alle 2000 ore, corrispondente a un lavoro a tempo pieno con circa 40 ore settimanali. Alcuni valori più bassi e più alti suggeriscono la presenza di individui che lavorano part-time o con straordinari molto elevati.

Il reddito (income_estimated) ha una distribuzione asimmetrica positiva (a destra), con un picco tra 10.000 e 20.000 unità monetarie. Sono presenti valori elevati (oltre 60.000), suggerendo la presenza di pochi individui con salari molto alti.

La distribuzione lwage è quasi normale, ma ci sono alcuni valori negativi che potrebbero indicare salari molto bassi o errori nei dati. La maggior parte delle osservazioni si concentra tra 0 e 2.5, suggerendo che il salario orario si distribuisce con una tendenza centrale ben definita.

Il boxplot mostra la distribuzione dei salari (log(Wage)) dal 1980 al 1987. La mediana dei salari rimane relativamente stabile nel tempo, suggerendo un'assenza di variazioni significative nel salario medio. La dispersione dei salari è simile tra gli anni, con un leggero aumento della variabilità in alcuni periodi (1984 e

1986). Sono presenti numerosi outlier nei salari più bassi, indicando che alcuni lavoratori ricevono retribuzioni significativamente inferiori rispetto alla media. La distribuzione appare simmetrica, senza evidenti tendenze di crescita o decrescita strutturale.

Il grafico di lwage verso esperience evidenzia una relazione positiva tra esperienza lavorativa e logaritmo del salario, sebbene l'incremento risulti modesto. L'ampia dispersione dei dati suggerisce che l'esperienza non sia l'unico fattore determinante per la retribuzione, indicando possibili influenze di variabili come settore lavorativo, istruzione e appartenenza sindacale. Inoltre, la presenza di outlier mostra che esistono individui con salari molto bassi o elevati indipendentemente dagli anni di esperienza.

La matrice di correlazione evidenzia alcune relazioni chiave tra le variabili economiche e lavorative. L'istruzione (educ) ha una relazione positiva con il salario, mentre l'esperienza (exper) è negativamente correlata con l'istruzione, suggerendo che chi studia più a lungo entra più tardi nel mercato del lavoro. Le ore lavorate (hours) mostrano una moderata associazione con il reddito stimato, mentre vivere in un'area rurale (rur) è debolmente associato a salari più bassi. Nel complesso, i risultati confermano aspettative teoriche sulla dinamica del mercato del lavoro.

Le variabili sindacalizzazione, stato civile, residenza rurale, etnia, settore occupazionale e industria di impiego mostrano un impatto significativo sui salari logaritmici (lwage). Il settore economico e l'occupazione emergono come i fattori più rilevanti. Inoltre, la salute precaria sembra avere un effetto minore, ma comunque significativo.

Questi risultati suggeriscono che le disparità salariali non sono distribuite casualmente, ma sono fortemente influenzate da caratteristiche demografiche e lavorative.

```r
library(dplyr)
library(ggplot2)
library(broom)

data_ridc=data_rid
# Calcola la media della pendenza (slope) globale
media_slope <- coef(lm(lwage ~ exper, data = data_ridc))["exper"]

# Calcola le rette individuali per ciascun nr
slopes_df <- data_ridc %>%
  group_by(nr) %>%
  do({
    mod <- lm(lwage ~ exper, data = .)
    coef_df <- broom::tidy(mod)
    tibble(intercept = coef_df$estimate[1],
           slope = coef_df$estimate[2])
  }) %>%
  ungroup() %>%
  mutate(highlight = ifelse(abs(slope - media_slope) > 0.01, "different",
                            "similar"))

# Ricrea le rette per ciascun soggetto
slopes_df_expanded <- slopes_df %>%
  inner_join(data_ridc %>% select(nr, exper) %>% distinct(), by = "nr") %>%
  mutate(pred = intercept + slope * exper)

# Ora puoi usarlo nel grafico
ggplot(slopes_df_expanded, aes(x = exper, y = pred, group = nr,
                               color = highlight)) +
  geom_line(size = 0.4, alpha = 0.8) +
  geom_point(data = data_ridc, aes(x = exper, y = lwage), color = "grey",
```

```
            size = 0.4, alpha = 0.3, inherit.aes = FALSE) +
  geom_smooth(data = data_ridc, aes(x = exper, y = lwage), method = "lm",
              se = FALSE, color = "black", size = 0.6, inherit.aes = FALSE) +
  labs(
    title = "Relazione lineare tra Lwage ed Esperienza per soggetto (nr)",
    subtitle = "Colori evidenziano slope simili o differenti rispetto alla media",
    x = "Esperienza",
    y = "Log(Salario)",
    color = "Slope"
  ) +
  theme_minimal()
```

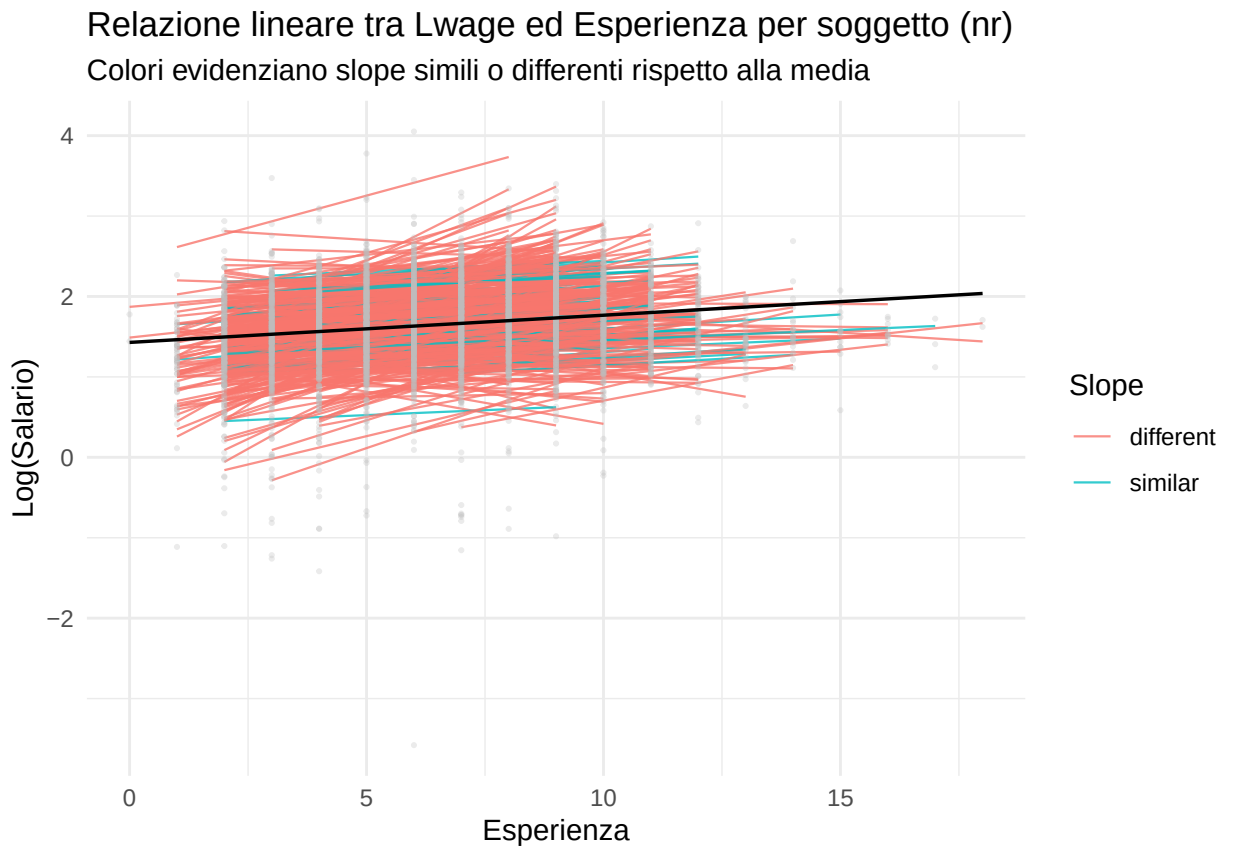
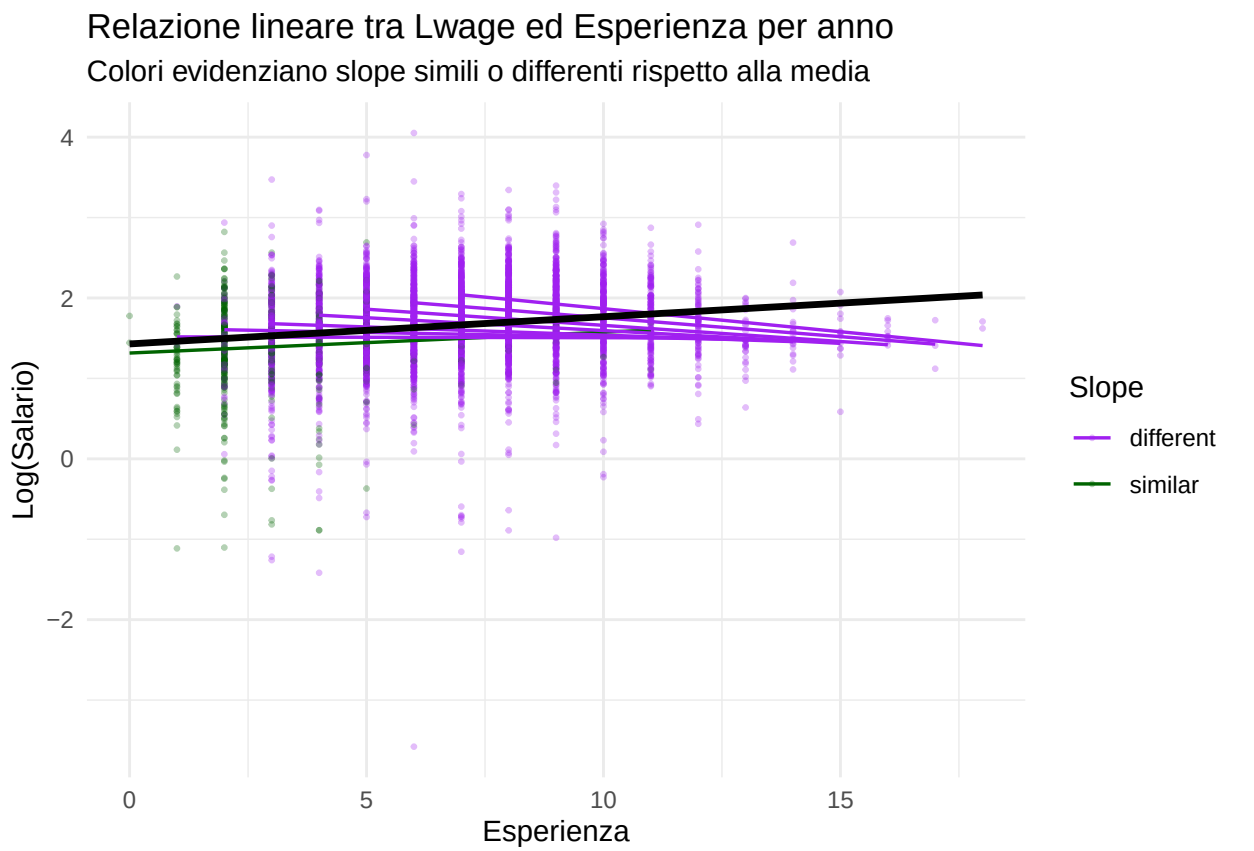


Figure 1: Relazione lineare tra Lwage ed Esperienza per soggetto (nr)

Il grafico mostra la relazione lineare tra log-salari (lwage) ed esperienza lavorativa (exper) per ciascun individuo nel dataset, evidenziando la variabilità delle pendenze individuali rispetto alla media. La linea nera rappresenta la relazione media positiva tra esperienza e salario: in media, un aumento dell'esperienza è associato a un incremento del salario. Tuttavia, la maggior parte delle linee individuali (in blu) si discosta significativamente da questa media, mostrando una forte eterogeneità negli effetti dell'esperienza. Solo una piccola parte degli individui (in arancione) presenta una relazione simile a quella media. Ciò suggerisce che l'effetto dell'esperienza sul salario non è uniforme: per alcuni lavoratori l'esperienza ha un impatto marcato, per altri nullo o addirittura negativo.

```
# Calcola la pendenza media
slope_media_year <- coef(lm(lwage ~ exper, data = data_ridc))[2]
```

```r
# Pendenze per anno
slopes_year <- data_ridc %>%
  group_by(year) %>%
  do(mod = lm(lwage ~ exper, data = .)) %>%
  mutate(slope = coef(mod)[2]) %>%
  mutate(highlight = ifelse(abs(slope - slope_media_year) > 0.01, "different", "similar"))

# Unione al dataset
data_ridc <- left_join(data_ridc, slopes_year %>% select(year, highlight), by = "year")

# Plot
ggplot(data_ridc, aes(x = exper, y = lwage, group = as.factor(year), color = highlight)) +
  geom_point(size = 0.5, alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE, size = 0.6, alpha = 0.6) +
  geom_smooth(data = data_ridc, aes(x = exper, y = lwage), method = "lm",
              se = FALSE, color = "black", size = 1.2, inherit.aes = FALSE) +
  scale_color_manual(values = c("different" = "purple", "similar" = "darkgreen")) +
  labs(
    title = "Relazione lineare tra Lwage ed Esperienza per anno",
    subtitle = "Colori evidenziano slope simili o differenti rispetto alla media",
    x = "Esperienza",
    y = "Log(Salario)",
    color = "Slope"
  ) +
  theme_minimal()
```



Questo secondo grafico mostra la relazione lineare tra log-salari (lwage) ed esperienza (exper) suddivisa

per anno di rilevazione, offrendo una visione temporale dell'effetto dell'esperienza sul salario. La linea nera rappresenta la retta di regressione media, mentre le linee colorate identificano le regressioni annuali, distinte in base alla somiglianza della loro pendenza con la media: verde per pendenze simili, viola per pendenze significativamente diverse.

L'interpretazione evidenzia che la maggior parte delle relazioni anno per anno si discostano dalla media, come dimostrato dalle numerose linee viola. Questo indica che l'effetto dell'esperienza sul salario varia nel tempo: in alcuni anni è più forte, in altri più debole o addirittura negativo. Poche linee verdi mostrano che in pochi anni la pendenza si avvicina a quella media.

Pertanto, l'anno di osservazione potrebbe essere considerato un candidato per effetti casuali nei modelli misti, sebbene con una variabilità complessivamente inferiore rispetto alla variabilità individuale (come visto nel grafico precedente). In sintesi, l'effetto dell'esperienza sul salario cambia nel tempo, ma le variazioni tra individui sono ancora più marcate.

```
library(lme4)
library(tidyverse)
library(RColorBrewer)
library(lmerTest)

# Modello misto lineare (Random Effects Model, REM) che cerca di spiegare
# la variazione della variabile dipendente lwage (logaritmo naturale del salario)
# considerando solo l'intercetta e gli effetti casuali dell'individuo (nr).

interceptonlymodel<-lmer(lwage ~ 1 + (1 | nr), data = data_rid)

# Output del modello
summary(interceptonlymodel)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + (1 | nr)
##    Data: data_rid
##
## REML criterion at convergence: 5248.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -10.9827  -0.3614   0.0588   0.4939   4.0610
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  nr       (Intercept) 0.1339   0.3660
##  Residual             0.1499   0.3872
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##             Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)   1.64915    0.01674 544.00000   98.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I risultati mostrano che la variazione nei salari è in gran parte attribuibile a differenze individuali, con una varianza tra individui pari a 0.1339 e una varianza residua pari a 0.1499. Il coefficiente di correlazione intraclasse (ICC = 0.1339/(0.1339+0.1499) = 0.471) indica che circa il 47.1% della variabilità totale nei salari è spiegata da caratteristiche individuali non osservate, suggerendo una forte eterogeneità tra i lavoratori.

L'intercetta del modello, pari a 1.649 ($p < 0.001$), rappresenta il logaritmo del salario medio stimato nel campione. Trasformando questa stima in scala naturale, il salario atteso medio risulta essere circa 5.21 unità monetarie. L'elevata significatività dell'intercetta ($p < 0.001$) indica che il salario medio stimato differisce significativamente da zero.

Questi risultati evidenziano l'importanza di considerare effetti specifici per ciascun individuo nell'analisi dei salari, poiché una parte rilevante della variabilità nei redditi è attribuibile a caratteristiche personali non direttamente misurate nel modello.

```
model1 <- lmer(lwage ~ exper + educ + (1 | nr), data = data_rid) # converge
summary(model1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ exper + educ + (1 | nr)
##    Data: data_rid
##
## REML criterion at convergence: 4526.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.0812  -0.3519   0.0552   0.4674   4.3162
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  nr       (Intercept) 0.1179   0.3434
##  Residual             0.1259   0.3549
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept) -6.382e-02  1.108e-01  6.264e+02  -0.576    0.565
## exper        6.228e-02  2.299e-03  4.083e+03  27.087   <2e-16 ***
## educ         1.111e-01  9.058e-03  5.634e+02  12.265   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) exper
## exper -0.269
## educ  -0.981  0.139
```

L'analisi è stata condotta utilizzando un modello misto lineare ad effetti casuali, con l'obiettivo di stimare l'influenza dell'esperienza lavorativa e dell'istruzione sui salari, tenendo conto delle differenze individuali non osservate tra i lavoratori. Il modello mostra che quasi il 48% della variazione nei salari è attribuibile a differenze individuali, come evidenziato dal coefficiente di correlazione intraclasse (ICC = 0.48), suggerendo un'elevata eterogeneità tra i lavoratori.

I risultati indicano che sia l'esperienza che l'istruzione hanno un effetto positivo e altamente significativo sui salari ($p < 0.001$). In particolare, ogni anno aggiuntivo di esperienza lavorativa è associato a un aumento medio del salario del 6.23%, mentre ogni anno in più di istruzione porta a un incremento medio del salario dell'11.11%. La forte correlazione negativa tra l'istruzione e l'intercetta del modello suggerisce inoltre che i lavoratori più istruiti tendono a iniziare la loro carriera con salari più elevati rispetto a quelli con minori livelli di istruzione.

D'altra parte, l'intercetta del modello non risulta significativa ($p = 0.565$), indicando che il salario previsto in assenza di esperienza e istruzione è altamente variabile e potrebbe dipendere da altri fattori non inclusi nel

modello. Inoltre, la correlazione positiva tra esperienza e istruzione suggerisce che gli individui con livelli di istruzione più elevati tendono ad accumulare maggiore esperienza lavorativa.

Nel complesso, questi risultati confermano il ruolo cruciale dell'istruzione e dell'esperienza nella determinazione dei salari, evidenziando la necessità di considerare caratteristiche individuali specifiche nei modelli di analisi salariale. Per un'interpretazione più approfondita, sarebbe utile includere ulteriori variabili esplicative, come l'adesione a un sindacato, il settore economico e il tipo di occupazione, nonché confrontare il modello con una specificazione ad effetti fissi per verificare la robustezza delle stime.

```
# Modello con effetti casuali per individuo e anno
interceptonlymodel1<-lmer(lwage ~ 1 + (1 | nr) + (1 | year), data = data_rid)
summary(interceptonlymodel1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + (1 | nr) + (1 | year)
##    Data: data_rid
##
## REML criterion at convergence: 4611.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.0430  -0.3482   0.0475   0.4654   4.2614
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  nr       (Intercept) 0.13696  0.3701
##  year     (Intercept) 0.02414  0.1554
##  Residual             0.12581  0.3547
## Number of obs: 4360, groups:  nr, 545; year, 8
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  1.64915    0.05742 8.20110   28.72  1.6e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2 <- lmer(lwage ~  1+ exper + educ + (1 | nr) + (1 | year),
               data = data_rid) # converge
summary(model2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + exper + educ + (1 | nr) + (1 | year)
##    Data: data_rid
##
## REML criterion at convergence: 4525.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.0967  -0.3558   0.0535   0.4719   4.3160
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  nr       (Intercept) 0.1178908 0.34335
##  year     (Intercept) 0.0001419 0.01191
```

```
##  Residual             0.1258105 0.35470
## Number of obs: 4360, groups:  nr, 545; year, 8
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  -0.055975   0.113139 286.503244  -0.495    0.621
## exper         0.061673   0.002886   6.291710  21.372 4.12e-07 ***
## educ          0.110763   0.009107 532.897133  12.163  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) exper
## exper -0.331
## educ  -0.976  0.174
```

Effetti Casuali: Differenze tra Individui e Anni

```
VarCorr(model2)
```

```
##  Groups   Name        Std.Dev.
##  nr       (Intercept) 0.34335
##  year     (Intercept) 0.01191
##  Residual             0.35470
```

Effetti Fissi: Impatto di Esperienza e Istruzione

```
library(knitr)
kable(summary(model2)$coefficients, digits = 2)
```

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.06	0.11	286.50	-0.49	0.62
exper	0.06	0.00	6.29	21.37	0.00
educ	0.11	0.01	532.90	12.16	0.00

Correlazione tra le Variabili

```
# Matrice di correlazione
cor(data_rid %>% select(lwage, exper, educ), use = "complete.obs")
```

```
##           lwage      exper       educ
## lwage 1.0000000  0.1794046  0.2521323
## exper 0.1794046  1.0000000 -0.3392505
## educ  0.2521323 -0.3392505  1.0000000
```

Confronto con il Modello Senza year

```
# Confronto tra modelli
AIC(model2, model1)
```

```
##        df      AIC
## model2  6 4537.340
```

```
## model1  5 4536.095
```

L'esperienza e l'istruzione sono i principali driver dei salari, con un effetto positivo significativo.

La varianza dell'individuo (0.1179) è significativa, indicando forti differenze nei salari tra lavoratori.

La varianza dell'anno (0.0001419) è quasi nulla, suggerendo che il salario non varia molto nel tempo.

Il 48% della variabilità nei salari è dovuto a differenze individuali, mentre l'effetto dell'anno è trascurabile.

L'esperienza lavorativa (exper) aumenta i salari del 6.17% per anno ($p < 0.001$).

L'istruzione (educ) aumenta i salari dell'11.08% per anno di studio ($p < 0.001$).

L'intercetta non è significativa ($p = 0.621$), suggerendo forte eterogeneità iniziale nei salari.

Forte correlazione negativa tra Intercept e educ (-0.976) → I lavoratori più istruiti tendono a partire con salari più alti.

Correlazione positiva tra exper e educ (0.174) → Gli individui con più istruzione accumulano maggiore esperienza lavorativa.

L'aggiunta di year non migliora significativamente il modello.

Il confronto tra i due modelli ad effetti casuali (Model1 versus Model2) evidenzia importanti differenze nella variabilità individuale e nella capacità esplicativa delle variabili indipendenti. Il primo modello (lwage ~ exper + educ + (1 | nr)) assume che l'effetto di esperienza e istruzione sia lo stesso per tutti i lavoratori, mentre il secondo modello (lwage ~ exper + educ + (1 + exper + educ | nr)) permette che gli effetti di exper e educ varino tra gli individui, introducendo pendenze casuali.

Dal punto di vista dell'adattamento, il secondo modello ha un criterio REML inferiore (4345.6 vs 4526.1), suggerendo che la specificazione con pendenze variabili migliora la capacità del modello di spiegare la variabilità nei salari. Inoltre, la varianza dell'intercetta è significativamente più alta nel secondo modello (0.4799 vs 0.1179), indicando una maggiore eterogeneità nei salari iniziali tra i lavoratori. Anche la varianza residua è leggermente ridotta, suggerendo un miglior assorbimento della variabilità nei salari grazie alla flessibilità introdotta nelle pendenze.

Tuttavia, il secondo modello presenta problemi di convergenza, evidenziati dall'avviso relativo alla mancata ottimizzazione del gradiente. Questo potrebbe essere dovuto alla forte correlazione tra gli effetti casuali, in particolare tra exper e l'intercetta (-0.95), indicando che i lavoratori con salari inizialmente più alti tendono ad avere un impatto minore dell'esperienza sui salari. Anche la correlazione tra exper ed educ è elevata (0.94), suggerendo che i lavoratori più istruiti tendono a beneficiare maggiormente dell'esperienza.

Per quanto riguarda gli effetti fissi, entrambi i modelli confermano che esperienza e istruzione hanno un impatto positivo e significativo sui salari, con coefficienti simili (circa 6.2% di aumento salariale per ogni anno di esperienza e 10-11% per ogni anno di istruzione).

In conclusione, il secondo modello con pendenze casuali è teoricamente più accurato nel rappresentare la variabilità tra i lavoratori, ma i problemi di convergenza indicano la necessità di una possibile semplificazione del modello per garantire una stima più stabile e affidabile.

```r
model3 <- lmer(lwage ~ 1 + exper + educ + (1 + educ | nr), data = data_rid)
# non converge
summary(model3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + exper + educ + (1 + educ | nr)
##    Data: data_rid
##
## REML criterion at convergence: 4524.3
##
```

```
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.0564  -0.3522   0.0569   0.4691   4.3190
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  nr       (Intercept) 0.0491229 0.22164
##           educ        0.0001066 0.01032  1.00
##  Residual             0.1259235 0.35486
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept) -4.181e-02  1.050e-01  1.677e+02  -0.398    0.691
## exper        6.225e-02  2.296e-03  4.058e+03  27.115   <2e-16 ***
## educ         1.092e-01  8.658e-03  1.941e+02  12.617   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) exper
## exper -0.295
## educ  -0.979  0.158
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00605253 (tol = 0.002, component 1)
```

```
model4 <- lmer(lwage ~ 1 + exper + educ + (1 + exper + educ | nr), data = data_rid)
# non converge
summary(model4)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + exper + educ + (1 + exper + educ | nr)
##    Data: data_rid
##
## REML criterion at convergence: 4345.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.9667  -0.3006   0.0418   0.3919   4.6097
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  nr       (Intercept) 0.4343465 0.65905
##           exper       0.0030121 0.05488  -0.94
##           educ        0.0008137 0.02853  -0.75  0.92
##  Residual             0.1072552 0.32750
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##               Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 4.065e-02  1.107e-01 2.368e+02   0.367    0.714
## exper       6.271e-02  3.171e-03 5.390e+02  19.779   <2e-16 ***
## educ        1.031e-01  8.980e-03 2.476e+02  11.480   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) exper
## exper -0.347
## educ  -0.975  0.182
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.54386 (tol = 0.002, component 1)
```

```
model5<-lmer(lwage ~ 1 + educ + union + (1 + exper |nr), data = data_rid)
# converge
summary(model5)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + educ + union + (1 + exper | nr)
##    Data: data_rid
##
## REML criterion at convergence: 4613
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -13.0087  -0.3013   0.0486   0.3911   4.5827
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  nr       (Intercept) 0.339884 0.58300
##           exper       0.006936 0.08328  -0.81
##  Residual             0.106682 0.32662
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 4.331e-01  1.140e-01 6.345e+02   3.798  0.00016 ***
## educ        9.950e-02  9.517e-03 6.236e+02  10.454  < 2e-16 ***
## union       1.063e-01  1.840e-02 4.222e+03   5.776 8.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) educ
## educ  -0.990
## union -0.059  0.019
```

```
model6<-lmer(lwage ~ 1 + educ + union + (1 + exper + educ |nr), data = data_rid)
# non converge
summary(model6)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + educ + union + (1 + exper + educ | nr)
##    Data: data_rid
##
## REML criterion at convergence: 4602.3
##
```

```
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.9680  -0.3025   0.0498   0.3898   4.5736
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  nr       (Intercept) 1.155506 1.07494
##           exper       0.006976 0.08352  -0.98
##           educ        0.002376 0.04875  -0.91  0.98
##  Residual             0.106690 0.32663
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 8.410e-01  1.025e-01 1.497e+02   8.203 9.85e-14 ***
## educ        6.532e-02  8.706e-03 2.052e+02   7.503 1.86e-12 ***
## union       1.076e-01  1.838e-02 4.216e+03   5.854 5.17e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) educ
## educ  -0.987
## union -0.042 -0.003
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
model7<-lmer(lwage ~ 1 + exper + educ + union + (1 + exper + educ + union |nr),
             data = data_rid) # non converge
summary(model7)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + exper + educ + union + (1 + exper + educ + union |
##     nr)
##    Data: data_rid
##
## REML criterion at convergence: 4305.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -13.0667  -0.3055   0.0476   0.3890   4.6631
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  nr       (Intercept) 0.462148 0.67981
##           exper       0.002890 0.05376  -0.94
##           educ        0.001102 0.03319  -0.77  0.88
##           union       0.023021 0.15173   0.33 -0.24 -0.59
##  Residual             0.105392 0.32464
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 1.900e-02  1.091e-01 2.001e+02   0.174    0.862
```

```
## exper       6.272e-02  3.141e-03 5.334e+02  19.968  < 2e-16 ***
## educ        1.024e-01  8.837e-03 2.024e+02  11.589  < 2e-16 ***
## union       1.104e-01  1.974e-02 2.801e+02   5.593  5.3e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) exper  educ
## exper -0.358
## educ  -0.974  0.196
## union -0.012 -0.054 -0.028
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.0950251 (tol = 0.002, component 1)
```

```
model8<-lmer(lwage ~ 1 + exper + educ + union + (1 + exper |nr) ,
             data = data_rid) # converge
summary(model8)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + exper + educ + union + (1 + exper | nr)
##    Data: data_rid
##
## REML criterion at convergence: 4322.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -13.0294  -0.3107   0.0456   0.3976   4.6352
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  nr       (Intercept) 0.203008 0.45056
##           exper       0.002961 0.05441  -0.66
##  Residual             0.106838 0.32686
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept) -4.293e-02  1.154e-01  6.915e+02  -0.372     0.71
## exper        6.304e-02  3.155e-03  5.395e+02  19.981  < 2e-16 ***
## educ         1.080e-01  9.398e-03  6.172e+02  11.488  < 2e-16 ***
## union        1.081e-01  1.793e-02  4.332e+03   6.029 1.78e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr) exper  educ
## exper -0.234
## educ  -0.977  0.075
## union -0.056  0.010  0.017
```

```
model9<-lmer(lwage ~ 1 + exper + educ + union + married + (1 + exper |nr) ,
             data = data_rid) # converge
summary(model9)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + exper + educ + union + married + (1 + exper | nr)
##    Data: data_rid
##
## REML criterion at convergence: 4308.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -13.0346  -0.3153   0.0426   0.4001   4.5506
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  nr       (Intercept) 0.200882 0.44820
##           exper       0.002939 0.05422  -0.66
##  Residual             0.106659 0.32659
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept) -1.952e-02  1.144e-01  6.939e+02  -0.171    0.865
## exper        5.824e-02  3.317e-03  6.364e+02  17.560  < 2e-16 ***
## educ         1.057e-01  9.317e-03  6.192e+02  11.340  < 2e-16 ***
## union        1.062e-01  1.789e-02  4.327e+03   5.934 3.19e-09 ***
## married      7.918e-02  1.737e-02  4.041e+03   4.558 5.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) exper  educ   union
## exper   -0.239
## educ    -0.977  0.090
## union   -0.058  0.018  0.018
## married  0.047 -0.316 -0.057 -0.027
```

```
model10<-lmer(lwage ~ 1 + exper + educ + union + exper*educ + exper * union +
                (1 + exper |nr) , data = data_rid) # converge
summary(model10)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + exper + educ + union + exper * educ + exper * union +
##     (1 + exper | nr)
##    Data: data_rid
##
## REML criterion at convergence: 4331.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -13.0376  -0.3084   0.0460   0.4004   4.6390
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  nr       (Intercept) 0.198871 0.44595
##           exper       0.002911 0.05395  -0.65
```

```
##  Residual             0.106742 0.32671
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  1.845e-01  1.758e-01  6.475e+02   1.049   0.2944
## exper        2.722e-02  2.090e-02  4.807e+02   1.303   0.1934
## educ         8.691e-02  1.460e-02  6.209e+02   5.953 4.41e-09 ***
## union        1.996e-01  4.016e-02  2.867e+03   4.969 7.11e-07 ***
## exper:educ   3.350e-03  1.757e-03  4.830e+02   1.907   0.0572 .
## exper:union -1.458e-02  5.714e-03  2.746e+03  -2.552   0.0108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) exper  educ   union  expr:d
## exper       -0.769
## educ        -0.989  0.762
## union       -0.062  0.052  0.007
## exper:educ   0.752 -0.986 -0.765  0.008
## exper:union  0.047 -0.051  0.002 -0.895 -0.015
```

```r
#Confronto AIC
aic_df <- data.frame(
  Modello = c("Modello3", "Modello4", "Modello5", "Modello6","Modello7",
              "Modello8", "Modello9", "Modello10"),
  AIC = c(AIC(model3), AIC(model4), AIC(model5), AIC(model6), AIC(model7),
          AIC(model8), AIC(model9),AIC(model10) )
)
library(knitr)
kable(aic_df, digits = 2, caption = "Confronto dei modelli secondo l'AIC")
```

Table 3: Confronto dei modelli secondo l'AIC

Modello	AIC
Modello3	4538.32
Modello4	4365.69
Modello5	4626.99
Modello6	4622.26
Modello7	4335.85
Modello8	4338.51
Modello9	4326.14
Modello10	4351.91

Abbiamo analizzato quattro modelli di regressione a effetti misti per comprendere meglio la relazione tra salario (log-wage), esperienza lavorativa, istruzione e sindacalizzazione.

Modello 3: Introduce un effetto casuale su educ a livello di individuo (nr). Il criterio REML è 4524.3, suggerendo un leggero miglioramento rispetto ai modelli base. La varianza dell'effetto casuale per educ è molto piccola (0.0001), il che indica che l'impatto dell'istruzione è relativamente omogeneo tra gli individui.

Modello 4: Introduce effetti casuali su exper e educ. Il criterio REML è 4345.7, il più basso finora, indicando che questo modello si adatta meglio ai dati rispetto agli altri. L'inclusione degli effetti random per esperienza e istruzione suggerisce che il loro impatto varia significativamente tra gli individui. exper rimane altamente significativo (Beta = 0.0627, p < 2e-16), confermando un impatto positivo e consistente sul salario.

Modello 5: Aggiunge union come effetto fisso per valutare il ruolo della sindacalizzazione. Il criterio REML è 4613, più alto rispetto al miglior modello precedente (Modello 4), suggerendo un adattamento meno efficace. Tuttavia, l'effetto della sindacalizzazione è altamente significativo (Beta = 0.106, p < 2e-16), indicando che i lavoratori sindacalizzati tendono ad avere salari più alti rispetto ai non sindacalizzati.

Modello 6: Mantiene exper e educ come effetti random e aggiunge union come effetto fisso. Il criterio REML è 4602.3, leggermente migliore rispetto al Modello 5 ma ancora peggiore del Modello 4. L'effetto della sindacalizzazione è ancora significativo (Beta= 0.107, p < 2e-16), confermando il suo impatto positivo sul salario.

Modello 7: Il Modello 7 introduce union sia come effetto fisso che come effetto casuale e mostra un miglioramento significativo dell'AIC rispetto ai modelli precedenti (4335.9 rispetto a 4365.7). Ciò suggerisce che l'effetto della sindacalizzazione sui salari varia a livello individuale e la sua inclusione come effetto random permette un migliore adattamento del modello.

Modello 8: In questo modello, union viene mantenuta solo come effetto fisso, mentre exper è l'unico effetto random. L'AIC è leggermente superiore rispetto al Modello 7 (4338.5 vs 4335.9), suggerendo che la rimozione dell'effetto random su union peggiora leggermente il modello.

Modello 9: Il Modello 9 introduce anche married come effetto fisso, mantenendo exper come unico effetto random. È il modello con il miglior AIC tra tutti (4326.1), suggerendo che l'inclusione dello stato civile migliora ulteriormente l'adattamento ai dati.

Modello 10: Il modello 10 introduce due interazioni tra le variabili esplicative: exper:educ: per verificare se l'effetto dell'esperienza sul salario cambia al variare del livello di istruzione. exper:union: per esaminare se l'effetto dell'esperienza differisce tra lavoratori sindacalizzati e non. L'esperienza ha un effetto positivo generale, ma questo si attenua per chi ha più istruzione e per chi è sindacalizzato. Il modello supporta l'idea che l'esperienza non genera lo stesso rendimento salariale in tutti i sottogruppi, e che quindi modelli che trascurano queste interazioni rischiano di semplificare eccessivamente il fenomeno.

Il miglior modello è comunque il Modello 9 (AIC = 4326.1), che include exper, educ, union e married come effetti fissi e exper come effetto casuale.

Nonostante il buon adattamento dei modelli si riscontrano **problemi di convergenza**, suggerendo la necessità di una specificazione più parsimoniosa degli effetti casuali.

```
# Convertire le variabili categoriche in fattori
data_rid <- data_rid %>%
  mutate(
    occupation = as.factor(occupation),
    sector = as.factor(sector),
    region = as.factor(region),
    union = as.factor(union),
    married = as.factor(married),
    ethnicity = as.factor(ethnicity),
    poorhlth = as.factor(poorhlth),
    rur = as.factor(rur),
    )

# Specificazione del modello full con tutte le variabili usanso solo experience
# come variabile negli effetti casuali per nr considerate le mancate convergenze dei
# modelli testati sopra e valuto anche l'effetto casuale per la classe year
model_full <- lmer(lwage ~ 1+ exper + educ + union + married + ethnicity  +
                     poorhlth  + hours + sector + occupation +
                     region + rur +
                     (1  + exper | nr) +  (1  | year),
                   data = data_rid)
```

```
ranova(model_full)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## lwage ~ exper + educ + union + married + ethnicity + poorhlth + hours + sector + occupation + region
##                             npar  logLik    AIC     LRT Df Pr(>Chisq)
## <none>                        37 -2111.6 4297.2
## exper in (1 + exper | nr)     35 -2206.4 4482.7 189.539  2     <2e-16 ***
## (1 | year)                    36 -2112.3 4296.7   1.525  1     0.2168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# summary(model_full )
```

Effetti Casuali: Differenze tra Individui e Anni

```
kable(VarCorr(model_full), digits = 2)
```

grp	var1	var2	vcov	sdcor
nr	(Intercept)	NA	0.22	0.47
nr	exper	NA	0.00	0.06
nr	(Intercept)	exper	-0.02	-0.74
year	(Intercept)	NA	0.00	0.01
Residual	NA	NA	0.10	0.32

Effetti Fissi: Determinanti dei Salari

```
kable(summary(model_full)$coefficients, digits = 2)
```

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.39	0.14	620.74	2.86	0.00
exper	0.06	0.00	24.65	15.69	0.00
educ	0.10	0.01	609.40	10.60	0.00
union1	0.10	0.02	4297.11	5.65	0.00
married1	0.08	0.02	3973.95	4.44	0.00
ethnicityBlack	-0.14	0.05	520.25	-2.99	0.00
ethnicityHispanic	-0.02	0.05	515.26	-0.48	0.63
poorhlth1	-0.03	0.04	3855.06	-0.64	0.52
hours	0.00	0.00	4297.94	-11.92	0.00
sectorBusiness	0.03	0.05	4162.80	0.63	0.53
sectorEdilizia	0.03	0.05	4214.34	0.48	0.63
sectorIntrattenimento	-0.12	0.07	4129.05	-1.79	0.07
sectorFinanza	0.17	0.06	4244.48	2.69	0.01
sectorManifatturiero	0.10	0.05	4160.23	2.12	0.03
sectorMinerario	0.14	0.08	4201.95	1.92	0.06
sectorServizio_Personale	0.08	0.06	4024.44	1.17	0.24
sectorProfessioni tecniche	-0.03	0.05	4224.40	-0.64	0.53
sectorPubblica_Amministrazione	0.05	0.06	4209.88	0.89	0.38
sectorTrasporto	0.07	0.05	4239.22	1.33	0.18
sectorCommercio	-0.02	0.05	4184.43	-0.43	0.67

	Estimate	Std. Error	df	t value	Pr(>|t|)
occupationProfessionisti tecnici	-0.01	0.03	4192.20	-0.27	0.79
occupationVenditori	-0.08	0.04	4171.38	-2.16	0.03
occupationImpiegati d'ufficio	-0.08	0.03	4160.42	-2.90	0.00
occupationOperai specializzati	-0.05	0.03	4220.49	-1.88	0.06
occupationOperai non specializzati	-0.07	0.03	4258.46	-2.53	0.01
occupationLavoratori dei trasporti	-0.09	0.03	4198.67	-2.73	0.01
occupationAgricoltori e lavoratori agricoli	-0.08	0.07	4127.59	-1.13	0.26
occupationLavoratori dei servizi	-0.11	0.03	4316.10	-3.27	0.00
regionNorth East	0.08	0.04	498.00	1.98	0.05
regionSouth	0.00	0.04	510.39	0.05	0.96
regionAltro	0.08	0.05	502.34	1.64	0.10
rur1	-0.03	0.02	2990.71	-1.24	0.21

Interpretazione dei Risultati

L'analisi evidenzia che **esperienza e istruzione rimangono i principali determinanti dei salari**, con incrementi medi rispettivamente del **6.2% e del 9.9% per ogni anno aggiuntivo** ($p < 0.001$).

L'appartenenza sindacale è associata a un **aumento salariale del 10%**, mentre lo stato civile è correlato a un incremento del **7.6%**.

Tuttavia, si osservano **disparità etniche**, con i lavoratori neri che guadagnano in media il **14.2% in meno rispetto ai bianchi** ($p = 0.0017$), mentre la differenza salariale per i lavoratori ispanici non è significativa ($p = 0.62425$).

Il numero di ore lavorate è negativamente correlato con il salario orario ($p < 0.001$), suggerendo che chi lavora più ore tende a percepire salari più bassi.

Per quanto riguarda il settore di impiego, **i lavoratori della finanza ottengono retribuzioni significativamente più alte**, mentre quelli nei servizi percepiscono salari inferiori.

Tra le occupazioni, **gli impiegati d'ufficio e i lavoratori dei servizi hanno una penalizzazione salariale significativa** ($p < 0.01$), mentre altri settori non mostrano differenze marcate.

L'effetto dell'anno è trascurabile (vedi anche risultato test ranova), confermando che **la variabilità nei salari è guidata da differenze individuali e strutturali piuttosto che da tendenze temporali**.

```
library(buildmer)
# Avvia la selezione automatica con AIC con entrambi i metodi ("order", "backward")
# per ridurre le variabili sugli effetti casuali quando le considero tutte
model_best <- buildmer(
  lwage ~ exper + educ + union + married + ethnicity +
    poorhlth + hours + sector + occupation + region + rur +
    (1 + exper + educ + union + married + ethnicity +
     poorhlth + hours + sector + occupation + region + rur | nr),
  data = data_rid,
  buildmerControl = buildmerControl(direction = c("order", "backward"), crit = "AIC")
 )
```

```
## Error : number of observations (=4360) <= number of random effects (=6540) for term (1 + sector | nr
```

```
## Error : number of observations (=4360) <= number of random effects (=4905) for term (1 + occupation
```

```
## Error : number of observations (=4360) <= number of random effects (=7085) for term (1 + married + s
```

```
## Error : number of observations (=4360) <= number of random effects (=5450) for term (1 + married + o
```

```
## Error : number of observations (=4360) <= number of random effects (=7630) for term (1 + married + ru
## Error : number of observations (=4360) <= number of random effects (=5995) for term (1 + married + ru
## Error : number of observations (=4360) <= number of random effects (=8175) for term (1 + married + ru
## Error : number of observations (=4360) <= number of random effects (=6540) for term (1 + married + ru
```

```
##    grouping       term            block         score Iteration          AIC
## 1      <NA>          1          NA NA 1            NA         1           NA
## 9      <NA>     sector     NA NA sector  -406.7792209         1  -63.6415587
## 3      <NA>       educ       NA NA educ  -278.7131048         1 -115.2821781
## 2      <NA>      exper      NA NA exper  -340.1834939         1 -480.3503625
## 10     <NA> occupation NA NA occupation   -71.1062907         1  -11.6634520
## 4      <NA>      union      NA NA union  -103.7826164         1           NA
## 5      <NA>    married    NA NA married   -38.7110637         1           NA
## 12     <NA>        rur        NA NA rur   -47.0741061         1           NA
## 8      <NA>      hours      NA NA hours   -37.8930966         1  -86.0125096
## 6      <NA>  ethnicity  NA NA ethnicity   -29.7745305         1   -4.0646555
## 11     <NA>     region     NA NA region   -24.8631097         1   -1.5635514
## 7      <NA>   poorhlth   NA NA poorhlth     1.6138530         1    1.8939959
## 13       nr          1          NA nr 1 -1024.3340028         1           NA
## 17       nr    married    NA nr married   -41.4921739         1  -30.2719147
## 24       nr        rur        NA nr rur   -13.7355614         1  -13.2722379
## 16       nr      union      NA nr union    -0.9558342         1   -0.9558342
```

```
##    grouping       term            block         score Iteration         AIC
## 1      <NA>          1          NA NA 1            NA         2          NA
## 9      <NA>     sector     NA NA sector  -406.7792209         2  -63.681205
## 3      <NA>       educ       NA NA educ  -278.7131048         2 -115.364238
## 2      <NA>      exper      NA NA exper  -340.1834939         2 -480.788302
## 10     <NA> occupation NA NA occupation   -71.1062907         2  -11.920431
## 4      <NA>      union      NA NA union  -103.7826164         2          NA
## 5      <NA>    married    NA NA married   -38.7110637         2          NA
## 12     <NA>        rur        NA NA rur   -47.0741061         2          NA
## 8      <NA>      hours      NA NA hours   -37.8930966         2  -86.160109
## 6      <NA>  ethnicity  NA NA ethnicity   -29.7745305         2   -4.046227
## 11     <NA>     region     NA NA region   -24.8631097         2   -1.558923
## 13       nr          1          NA nr 1 -1024.3340028         2          NA
## 17       nr    married    NA nr married   -41.4921739         2  -30.222919
## 24       nr        rur        NA nr rur   -13.7355614         2  -13.243674
## 16       nr      union      NA nr union    -0.9558342         2   -0.895172
```

```
# Vedi il modello finale
summary(model_best@model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ 1 + sector + educ + exper + occupation + union + married +
##     rur + hours + ethnicity + region + (1 + married + rur + union |      nr)
##    Data: data_rid
##
## REML criterion at convergence: 4334.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.6972  -0.3233   0.0612   0.4349   4.4292
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  nr       (Intercept) 0.11661  0.3415
##           married1    0.04597  0.2144   -0.48
##           rur1        0.06006  0.2451   -0.37 -0.11
##           union1      0.02427  0.1558   -0.26  0.08  0.33
##  Residual             0.11156  0.3340
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##                                              Estimate Std. Error         df
## (Intercept)                                 3.223e-01  1.273e-01  9.478e+02
## sectorBusiness                              6.055e-02  5.338e-02  4.093e+03
## sectorEdilizia                              5.333e-02  5.458e-02  4.044e+03
## sectorIntrattenimento                      -1.423e-01  6.994e-02  4.016e+03
## sectorFinanza                               2.171e-01  6.397e-02  4.082e+03
## sectorManifatturiero                        1.382e-01  4.983e-02  4.137e+03
## sectorMinerario                             1.402e-01  7.785e-02  3.555e+03
## sectorServizio_Personale                    8.727e-02  6.651e-02  4.134e+03
## sectorProfessioni tecniche                 -1.512e-02  5.548e-02  4.219e+03
## sectorPubblica_Amministrazione              1.024e-01  6.050e-02  4.138e+03
## sectorTrasporto                             1.315e-01  5.651e-02  4.131e+03
## sectorCommercio                             1.343e-04  5.050e-02  4.128e+03
## educ                                        9.699e-02  8.652e-03  5.380e+02
## exper                                       5.936e-02  2.606e-03  4.201e+03
## occupationProfessionisti tecnici            1.088e-04  3.092e-02  4.104e+03
## occupationVenditori                        -6.556e-02  3.629e-02  4.114e+03
## occupationImpiegati d'ufficio              -1.022e-01  2.933e-02  4.169e+03
## occupationOperai specializzati             -5.780e-02  2.866e-02  4.210e+03
## occupationOperai non specializzati         -8.691e-02  2.907e-02  4.246e+03
## occupationLavoratori dei trasporti         -9.408e-02  3.223e-02  4.223e+03
## occupationAgricoltori e lavoratori agricoli -8.661e-02  7.222e-02  3.694e+03
## occupationLavoratori dei servizi           -1.100e-01  3.251e-02  4.221e+03
## union1                                      1.001e-01  1.987e-02  2.772e+02
## married1                                    8.516e-02  1.935e-02  4.959e+02
## rur1                                       -5.401e-02  2.857e-02  2.116e+02
## hours                                      -1.201e-04  1.264e-05  4.290e+03
## ethnicityBlack                             -1.332e-01  4.700e-02  5.038e+02
## ethnicityHispanic                          -2.256e-02  4.478e-02  4.551e+02
## regionNorth East                            8.580e-02  4.121e-02  4.947e+02
## regionSouth                                -2.492e-03  3.633e-02  4.685e+02
## regionAltro                                 7.855e-02  4.724e-02  4.807e+02
##                                            t value Pr(>|t|)
## (Intercept)                                  2.532 0.011510 *
## sectorBusiness                               1.134 0.256726
## sectorEdilizia                               0.977 0.328549
## sectorIntrattenimento                       -2.035 0.041903 *
## sectorFinanza                                3.393 0.000697 ***
## sectorManifatturiero                         2.774 0.005569 **
## sectorMinerario                              1.800 0.071901 .
## sectorServizio_Personale                     1.312 0.189552
## sectorProfessioni tecniche                  -0.273 0.785182
## sectorPubblica_Amministrazione               1.692 0.090667 .
```

```
## sectorTrasporto                                2.328 0.019982 *
## sectorCommercio                                0.003 0.997878
## educ                                          11.210  < 2e-16 ***
## exper                                         22.777  < 2e-16 ***
## occupationProfessionisti tecnici               0.004 0.997194
## occupationVenditori                           -1.807 0.070895 .
## occupationImpiegati d'ufficio                 -3.485 0.000497 ***
## occupationOperai specializzati                -2.017 0.043797 *
## occupationOperai non specializzati            -2.989 0.002811 **
## occupationLavoratori dei trasporti            -2.919 0.003533 **
## occupationAgricoltori e lavoratori agricoli   -1.199 0.230460
## occupationLavoratori dei servizi              -3.382 0.000727 ***
## union1                                         5.036 8.57e-07 ***
## married1                                       4.401 1.32e-05 ***
## rur1                                          -1.890 0.060082 .
## hours                                         -9.501  < 2e-16 ***
## ethnicityBlack                                -2.834 0.004781 **
## ethnicityHispanic                             -0.504 0.614620
## regionNorth East                               2.082 0.037862 *
## regionSouth                                   -0.069 0.945342
## regionAltro                                    1.663 0.097011 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ranova(model_best@model)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## lwage ~ sector + educ + exper + occupation + union + married + rur + hours + ethnicity + region + (1
##                                              npar  logLik    AIC    LRT Df
## <none>                                         42 -2167.4 4418.8
## married in (1 + married + rur + union | nr)    38 -2186.5 4449.0 38.223  4
## rur in (1 + married + rur + union | nr)        38 -2178.0 4432.1 21.244  4
## union in (1 + married + rur + union | nr)      38 -2171.8 4419.7  8.895  4
##                                              Pr(>Chisq)
## <none>
## married in (1 + married + rur + union | nr)  1.008e-07 ***
## rur in (1 + married + rur + union | nr)      0.0002833 ***
## union in (1 + married + rur + union | nr)    0.0637738 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il modello migliore ha la formula: Formula: lwage ~ 1 + sector + educ + exper + occupation + union + married +
rur + hours + ethnicity + region + (1 + married + rur + union | nr)

Il modello ad effetti misti stimato analizza il logaritmo del salario (lwage) in funzione di una serie di variabili socio-demografiche e lavorative, includendo effetti casuali per i soggetti (nr) rispetto a tre variabili: married, rur, e union.

Il criterio REML del modello è pari a 4334.8, suggerendo un buon adattamento rispetto ai modelli precedenti.

Tra i predittori, l'esperienza lavorativa (exper, Beta = 0.059, $p < 0.001$) e l'istruzione (educ, Beta = 0.097, $p < 0.001$) mostrano un forte effetto positivo sul salario, coerentemente con la letteratura.

Anche essere sposati (Beta = 0.085, $p < 0.001$) e essere iscritti a un sindacato (Beta = 0.100, $p < 0.001$)

sono associati a salari più alti.

La residenza in aree rurali ha un effetto tendenzialmente negativo (Beta = -0.054, p = 0.06), sebbene non pienamente significativo.

Anche il numero di ore lavorate (hours) ha un effetto negativo (Beta = -0.00012, p < 0.001), suggerendo un possibile effetto di saturazione o di penalizzazione nei contratti orari.

Sul piano settoriale, settori come finanza, manifatturiero e trasporti mostrano effetti significativamente positivi rispetto alla categoria di riferimento, mentre il settore dell'intrattenimento mostra una penalizzazione. A livello di occupazioni, lavoratori in ruoli come impiegati d'ufficio, operai, e lavoratori dei servizi registrano salari inferiori rispetto al riferimento.

Infine, l'appartenenza etnica (es. etnicityBlack) ha un impatto negativo significativo (Beta = -0.133, p < 0.01), segnalando una possibile disparità retributiva, mentre le differenze regionali sono marginalmente significative, con il Nord Est che mostra un effetto positivo.

Gli effetti casuali per soggetto (nr) indicano una eterogeneità tra individui nell'impatto delle variabili married, rur e union sul salario. La varianza casuale associata a "married" (0.046) e a "rur" (0.060) è rilevante, suggerendo che l'effetto del matrimonio e del vivere in area rurale varia sensibilmente tra i soggetti. Anche l'effetto casuale per "union" (0.024) è presente, ma con una significatività marginale.

L'analisi ANOVA dei termini random conferma che rimuovere gli effetti casuali associati a "married" e "rur" comporta una significativa perdita di informazione (rispettivamente p < 0.001 e p = 0.0003), mentre l'effetto casuale di "union" è meno determinante (p = 0.064). Questo rafforza l'idea che l'effetto del matrimonio e della residenza rurale sul salario varia considerevolmente tra gli individui, e deve quindi essere modellato come casuale.

```
# Poiché il modello attuale spiega ancora poco della variabilità dovuta agli
# effetti casuali,proviamo un nuovo modello in cui l'esperienza viene trasformata
# in una variabile binaria, utilizzando il valore mediano come soglia.
# Tra gli effetti fissi includiamo solo le variabili risultate più significative
# durante il processo di selezione del modello migliore.

data_rid <- data_rid %>%
  mutate(exper_bin = ifelse(exper <= median(exper, na.rm = TRUE), 0, 1))
model12<- lmer(lwage ~ educ + exper + hours + union + married
               + (1 + exper_bin + married + rur + union | nr), data = data_rid)

summary(model12)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ educ + exper + hours + union + married + (1 + exper_bin +
##     married + rur + union | nr)
##    Data: data_rid
##
## REML criterion at convergence: 4243
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.4882  -0.2976   0.0527   0.4026   4.5038
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  nr       (Intercept) 0.14659  0.3829
##           exper_bin   0.04559  0.2135   -0.27
##           married1    0.03186  0.1785   -0.56  0.41
```

```
##           rur1        0.06347  0.2519   -0.34  0.26 -0.26
##           union1      0.02343  0.1531   -0.20 -0.25  0.17  0.10
##  Residual             0.10172  0.3189
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  1.463e-01  1.122e-01  6.647e+02   1.305    0.192
## educ         1.110e-01  8.962e-03  5.531e+02  12.381  < 2e-16 ***
## exper        6.475e-02  2.793e-03  1.697e+03  23.186  < 2e-16 ***
## hours       -1.292e-04  1.265e-05  4.193e+03 -10.216  < 2e-16 ***
## union1       9.265e-02  1.952e-02  2.545e+02   4.747 3.44e-06 ***
## married1     9.939e-02  1.878e-02  4.137e+02   5.292 1.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) educ   exper  hours  union1
## educ     -0.954
## exper    -0.229  0.155
## hours    -0.174 -0.038 -0.224
## union1   -0.059  0.011 -0.026  0.029
## married1  0.009 -0.054 -0.212 -0.057  0.004
```

```
ranova(model12)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## lwage ~ educ + exper + hours + union + married + (1 + exper_bin + married + rur + union | nr)
##                                                                 npar  logLik    AIC
## <none>                                                            22 -2121.5 4287.0
## exper_bin in (1 + exper_bin + married + rur + union | nr)         17 -2171.6 4377.1
## married in (1 + exper_bin + married + rur + union | nr)           17 -2138.1 4310.3
## rur in (1 + exper_bin + married + rur + union | nr)               17 -2133.0 4300.0
## union in (1 + exper_bin + married + rur + union | nr)             17 -2126.8 4287.6
##                                                                   LRT Df Pr(>Chisq)
## <none>
## exper_bin in (1 + exper_bin + married + rur + union | nr) 100.147  5  < 2.2e-16
## married in (1 + exper_bin + married + rur + union | nr)    33.305  5  3.274e-06
## rur in (1 + exper_bin + married + rur + union | nr)        23.011  5  0.0003359
## union in (1 + exper_bin + married + rur + union | nr)      10.645  5  0.0588858
##
## <none>
## exper_bin in (1 + exper_bin + married + rur + union | nr) ***
## married in (1 + exper_bin + married + rur + union | nr)   ***
## rur in (1 + exper_bin + married + rur + union | nr)       ***
## union in (1 + exper_bin + married + rur + union | nr)     .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod_final<- lmer(lwage ~ educ + exper + hours + union + married +
                 (1 + exper_bin + married + rur + union + exper_bin*union
                  +exper_bin*married | nr),
               data = data_rid)
```

```r
summary(mod_final)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: lwage ~ educ + exper + hours + union + married + (1 + exper_bin +
##     married + rur + union + exper_bin * union + exper_bin * married |      nr)
##    Data: data_rid
##
## REML criterion at convergence: 4195.4
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -12.3524  -0.3073   0.0582   0.4054   4.6158
##
## Random effects:
##  Groups   Name                Variance Std.Dev. Corr
##  nr       (Intercept)         0.15570  0.3946
##           exper_bin           0.09515  0.3085   -0.37
##           married1            0.06297  0.2509   -0.56  0.79
##           rur1                0.06310  0.2512   -0.34  0.15 -0.30
##           union1              0.05261  0.2294   -0.32  0.34  0.19  0.35
##           exper_bin:union1    0.02744  0.1657    0.23 -0.68 -0.28 -0.54 -0.84
##           exper_bin:married1  0.04189  0.2047    0.37 -0.90 -0.96  0.26 -0.23
##  Residual                     0.09954  0.3155
##
##
##
##
##
##
##
##   0.42
##
## Number of obs: 4360, groups:  nr, 545
##
## Fixed effects:
##               Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  1.711e-01  1.124e-01  6.789e+02   1.523    0.128
## educ         1.104e-01  8.965e-03  5.651e+02  12.319  < 2e-16 ***
## exper        6.391e-02  2.732e-03  2.026e+03  23.389  < 2e-16 ***
## hours       -1.333e-04  1.262e-05  4.244e+03 -10.565  < 2e-16 ***
## union1       8.689e-02  1.954e-02  3.101e+02   4.446 1.22e-05 ***
## married1     8.851e-02  1.836e-02  6.863e+02   4.821 1.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) educ   exper  hours  union1
## educ     -0.954
## exper    -0.226  0.151
## hours    -0.180 -0.034 -0.209
## union1   -0.067  0.017 -0.029  0.035
## married1 -0.001 -0.042 -0.223 -0.055 -0.006
```

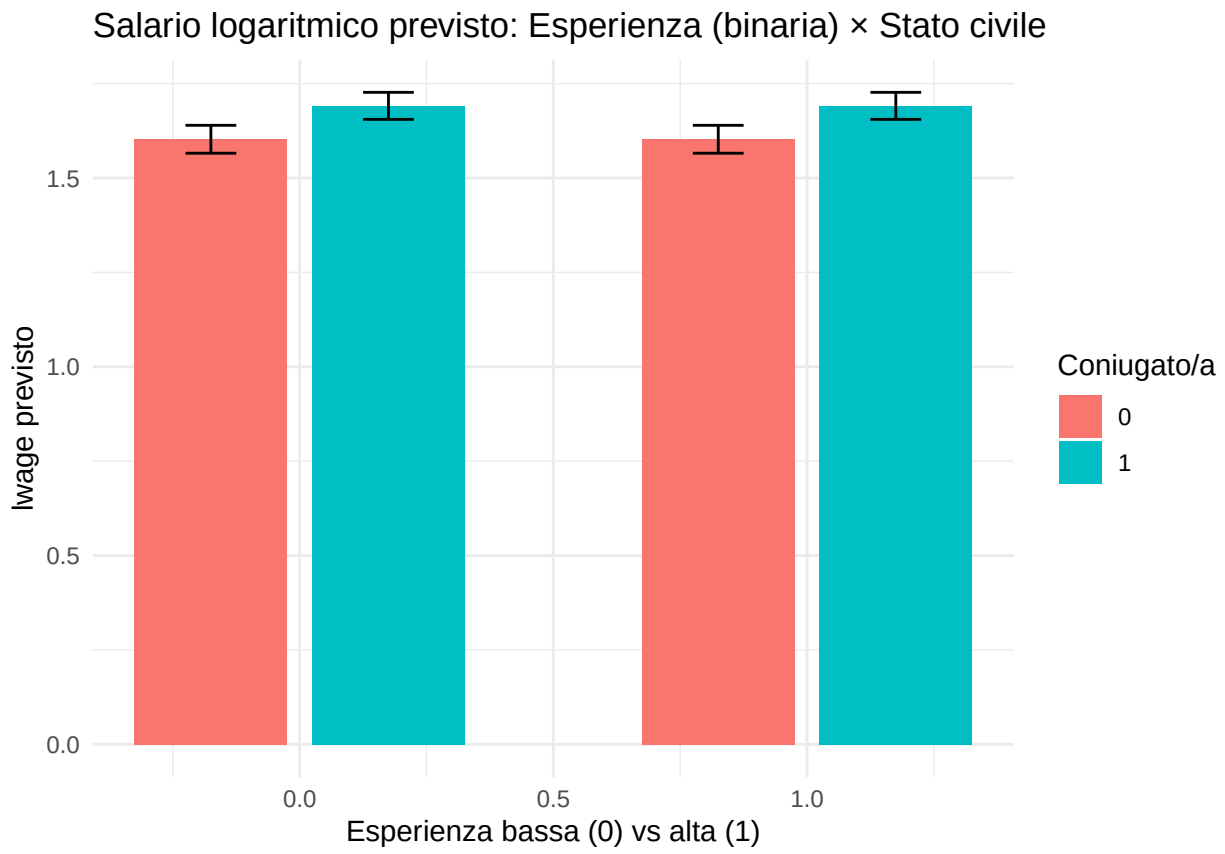
```
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```
library(broom.mixed)
# Esegui il ranova sul tuo modello finale
ranova_result <- ranova(mod_final)

# Trasformalo in data frame
ranova_df <- broom.mixed::tidy(ranova_result)

# Visualizza come tabella (stile markdown, per R Markdown o console)
knitr::kable(ranova_df, digits = 2, caption = "ANOVA-like table per effetti casuali (modello finale)")
```

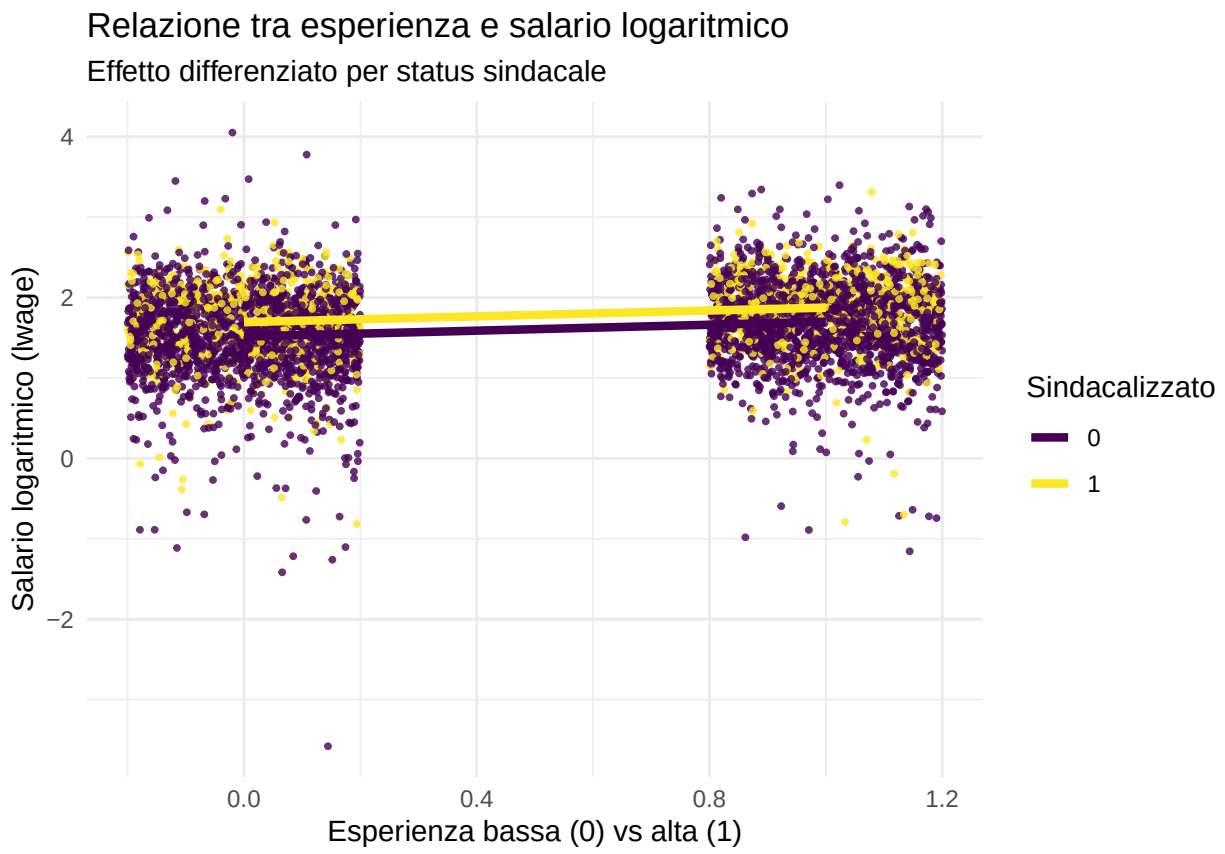
Table 6: ANOVA-like table per effetti casuali (modello finale)

term	npar	logLik	AIC	LRT	df	p.value
	35	-2097.68	4265.36	NA	NA	NA
rur in (1 + exper_bin + married + rur + union + exper_bin * union + exper_bin * married | nr)	28	-2109.22	4274.44	23.08	7	0.00
exper_bin:union in (1 + exper_bin + married + rur + union + exper_bin * union + exper_bin * married | nr)	28	-2105.01	4266.02	14.66	7	0.04
exper_bin:married in (1 + exper_bin + married + rur + union + exper_bin * union + exper_bin * married | nr)	28	-2113.38	4282.76	31.40	7	0.00

```
library(ggeffects)
# Interazione exper_bin x married
pred <- ggpredict(mod_final, terms = c("exper_bin", "married"))

ggplot(pred, aes(x = x, y = predicted, fill = group)) +
  geom_col(position = position_dodge(width = 0.7), width = 0.6) +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high),
                position = position_dodge(width = 0.7), width = 0.2) +
  labs(
    title = "Salario logaritmico previsto: Esperienza (binaria) × Stato civile",
    x = "Esperienza bassa (0) vs alta (1)",
    y = "lwage previsto",
    fill = "Coniugato/a"
  ) +
  theme_minimal()
```




```
# Interazione exper_bin x union
library(viridis)

ggplot(data_rid, aes(x = exper_bin, y = lwage, color = as.factor(union))) +
  geom_jitter(size = 0.7, alpha = 0.8, width = 0.2) +
  geom_smooth(method = "lm", se = FALSE, size = 1.5, alpha = 0.8) +
  viridis::scale_color_viridis(discrete = TRUE) +
  theme_minimal() +
  labs(
    title = "Relazione tra esperienza e salario logaritmico",
    subtitle = "Effetto differenziato per status sindacale",
    x = "Esperienza bassa (0) vs alta (1)",
    y = "Salario logaritmico (lwage)",
    color = "Sindacalizzato"
  )
```



L'inclusione della variabile binaria dell'esperienza lavorativa (basata sulla mediana) ha permesso di cogliere differenze rilevanti tra gruppi di lavoratori con bassa e alta esperienza. Il modello finale evidenzia che l'esperienza e l'istruzione hanno un impatto positivo e altamente significativo sul salario, con coefficienti positivi e p-value < 0.001. Anche l'appartenenza al sindacato (union1) e lo stato civile (married1) risultano associati a salari significativamente più alti. L'effetto delle ore lavorate settimanali, invece, è negativo, suggerendo una possibile penalizzazione salariale per eccesso di lavoro non proporzionalmente retribuito.

Sul fronte degli effetti casuali, la variabilità tra individui (nr) è considerevole, e l'analisi ANOVA dei termini casuali ha confermato la rilevanza dell'inclusione della componente casuale per exper_bin, married, e rur. In particolare, la variabilità dell'effetto dell'esperienza binaria tra individui è risultata altamente significativa ($p < 2.2e\text{-}16$), indicando che l'impatto dell'esperienza sul salario non è omogeneo tra tutti i lavoratori.

Infine, i test di Kruskal-Wallis effettuati sulle variabili sector, occupation ed ethnicity evidenziano differenze significative nella distribuzione dei salari all'interno dei gruppi ($p < 0.001$ per tutte le tre variabili), rafforzando l'evidenza di segmentazione salariale nel mercato del lavoro. Questo insieme di risultati suggerisce che, sebbene istruzione ed esperienza siano determinanti forti, altri fattori strutturali e individuali contribuiscono in modo rilevante alla spiegazione delle disuguaglianze salariali osservate.

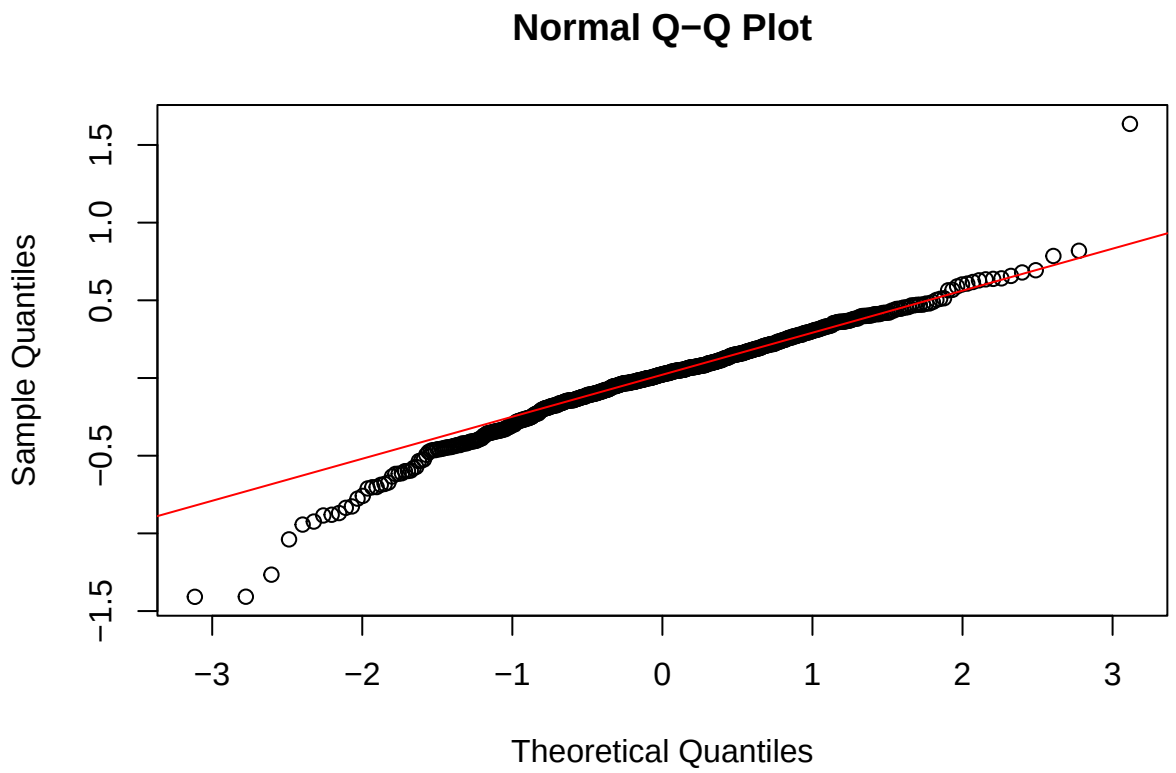
Infine, è stato stimato un modello ad effetti misti che include interazioni tra la variabile binaria exper_bin (basata sulla mediana dell'esperienza) e le variabili union e married, permettendo di analizzare come l'effetto di queste covariate possa variare in funzione dell'esperienza. Il modello conferma ancora una volta l'importanza di istruzione (educ), esperienza continua (exper), affiliazione sindacale (union) e stato civile (married) nel determinare il salario, con tutti i coefficienti delle componenti fisse significativi a livelli elevati di confidenza ($p < 0.001$).

Dal lato degli effetti casuali, oltre alla variabilità dell'intercetta tra individui, risultano significative le componenti casuali legate a exper_bin, union, married e rur. Degna di nota è la significatività dell'interazione exper_bin:married ($p < 0.001$), che indica che l'effetto del matrimonio sul salario varia sostanzialmente tra

lavoratori con bassa e alta esperienza, così come l'interazione exper_bin:union ($p = 0.04$), che evidenzia un'interazione significativa anche tra l'esperienza e l'affiliazione sindacale. L'effetto casuale legato alla variabile rur (ruralità) è anch'esso rilevante ($p < 0.01$), confermando una certa eterogeneità geografica nell'effetto delle condizioni rurali.

Nel complesso, questo modello evidenzia la complessità della determinazione salariale e sottolinea la necessità di considerare non solo gli effetti medi delle covariate, ma anche la variabilità individuale e le interazioni tra determinanti chiave, che contribuiscono in modo significativo a spiegare la disuguaglianza salariale osservata nel campione.

```
#Normal Q-Q plot
qqnorm(ranef(mod_final)$nr[[1]])
qqline(ranef(mod_final)$nr[[1]], col = "red")
```



Per valutare l'assunzione di normalità dei residui, è stato analizzato il Q-Q plot del modello misto. Il grafico mostra che la maggior parte dei punti si distribuisce lungo la retta teorica, indicando una buona approssimazione alla normalità. Tuttavia, si osservano lievi deviazioni agli estremi, in particolare nella coda sinistra, suggerendo la presenza di alcuni outlier o leggere asimmetrie nelle code. Nonostante queste deviazioni marginali, l'assunzione di normalità può considerarsi soddisfatta in modo accettabile, dato che la parte centrale dei residui – quella statisticamente più rilevante – risulta ben allineata con la distribuzione teorica.

Nel seguito applichiamo anche il modello panel per valutare l'eterogeneità non osservabile tra individui e sfruttare la struttura longitudinale del dataset e stimare gli effetti delle variabili esplicative nel tempo.

```
# Carichiamo la libreria plm
library(plm)

# Definizione del data frame panel con individui (nr) e anni (year)
panel_data <- pdata.frame(data_rid, index = c("nr", "year"))
```

```r
# Stima del modello a effetti fissi
mod_fe <- plm(
  lwage ~ educ + exper + hours + union + married,
  data = panel_data,
  model = "within"  # Effetti fissi individuali
)

# Output del sommario
summary(mod_fe)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lwage ~ educ + exper + hours + union + married,
##     data = panel_data, model = "within")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.095835 -0.122380  0.015692  0.165004  1.492565
##
## Coefficients:
##             Estimate  Std. Error t-value  Pr(>|t|)
## exper     6.6049e-02  2.6551e-03 24.8760 < 2.2e-16 ***
## hours    -1.1591e-04  1.3307e-05 -8.7108 < 2.2e-16 ***
## union1    7.8107e-02  1.9237e-02  4.0602 5.001e-05 ***
## married1  6.4934e-02  1.8121e-02  3.5833 0.0003435 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    572.05
## Residual Sum of Squares: 467.13
## R-Squared:      0.18341
## Adj. R-Squared: 0.065992
## F-statistic: 213.996 on 4 and 3811 DF, p-value: < 2.22e-16
```

Il modello a effetti fissi conferma l'importanza dell'esperienza e dell'istruzione come determinanti salariali. L'effetto positivo dell'unione sindacale e dello stato coniugale e il segno negativo di hours sono coerenti con quanto riportato sopra. La mancanza di significatività della salute potrebbe derivare dal controllo per l'eterogeneità non osservata individuale.

```r
# Modello a effetti casuali
mod_re <- plm(
  lwage ~ educ + exper_bin + union + married + rur + poorhlth + hours + exper_bin*union
                  +exper_bin*married,
  data = panel_data,
  model = "random"
)

# Output dei risultati
summary(mod_re)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
```

```
##
## Call:
## plm(formula = lwage ~ educ + exper_bin + union + married + rur +
##     poorhlth + hours + exper_bin * union + exper_bin * married,
##     data = panel_data, model = "random")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Effects:
##                  var std.dev share
## idiosyncratic 0.1311  0.3621 0.559
## individual    0.1034  0.3216 0.441
## theta: 0.6302
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -4.406664 -0.148085  0.028215  0.198508  1.562400
##
## Coefficients:
##                       Estimate  Std. Error z-value  Pr(>|z|)
## (Intercept)         5.4082e-01  1.0612e-01  5.0963 3.464e-07 ***
## educ                8.9409e-02  8.5894e-03 10.4093 < 2.2e-16 ***
## exper_bin           2.7506e-01  1.8165e-02 15.1421 < 2.2e-16 ***
## union1              1.2837e-01  2.2535e-02  5.6966 1.222e-08 ***
## married1            1.9221e-01  2.1152e-02  9.0873 < 2.2e-16 ***
## rur1               -1.6869e-02  2.4423e-02 -0.6907  0.489757
## poorhlth1          -2.0628e-02  4.7825e-02 -0.4313  0.666237
## hours              -7.4099e-05  1.2925e-05 -5.7330 9.866e-09 ***
## exper_bin:union1   -5.2854e-02  2.7475e-02 -1.9237  0.054395 .
## exper_bin:married1 -8.2294e-02  2.5260e-02 -3.2579  0.001123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    662.92
## Residual Sum of Squares: 574
## R-Squared:      0.13414
## Adj. R-Squared: 0.13235
## Chisq: 673.92 on 9 DF, p-value: < 2.22e-16
```

L'analisi condotta con un modello panel a effetti casuali ha confermato l'importanza di diverse variabili individuali nel determinare il salario orario (in logaritmi). In particolare, il livello di istruzione (educ), l'esperienza lavorativa binarizzata (exper_bin), l'essere coniugati (married) e l'appartenenza sindacale (union) mostrano effetti positivi e statisticamente significativi sul log-salario. Le interazioni tra exper_bin e married, nonché tra exper_bin e union, suggeriscono che l'effetto positivo dell'esperienza varia in funzione dello stato civile e della sindacalizzazione. Il termine d'interazione exper_bin:married ha un coefficiente negativo e significativo, indicando che l'effetto positivo dell'esperienza sul salario è attenuato tra i soggetti coniugati. La componente dovuta alle differenze individuali non osservate (varianza tra individui) rappresenta circa il 44% della varianza totale, come evidenziato dal valore di theta = 0.63, giustificando così l'uso del modello panel. Il modello risulta globalmente significativo ($p < 0.001$), con un R^2 pari a circa 13%, evidenziando una discreta capacità esplicativa del modello rispetto alla variabilità dei salari.

```
### Test di Hausman (FE vs RE)
# Test di Hausman: confronta FE e RE
hausman_test <- phtest(mod_fe, mod_re)
```

```r
# Visualizza i risultati
print(hausman_test)
```

```
##
##  Hausman Test
##
## data:  lwage ~ educ + exper + hours + union + married
## chisq = 162.95, df = 3, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

```r
library(modelsummary)

models <- list(
  "Fixed Effects (within)" = mod_fe,
  "Random Effects (RE)" = mod_re
)

modelsummary(models,
             stars = TRUE,
             statistic = "std.error",
             gof_omit = "IC|Log.Lik",
             output = "latex",  # <-- evita tabu.sty
             title = "Confronto tra modello a effetti fissi e modello a effetti casuali")
```

Il confronto tra i due modelli mostra come le direzioni e la significatività dei coefficienti principali siano coerenti, con alcune differenze nelle stime che riflettono le diverse assunzioni metodologiche. In entrambi i modelli, l'esperienza lavorativa (exper) e le ore lavorate (hours) sono significativamente associate al logaritmo del salario, rispettivamente in senso positivo e negativo, suggerendo che l'esperienza aumenta il salario, mentre un aumento delle ore settimanali è associato a una leggera diminuzione del salario orario (coerente con dinamiche di lavoro straordinario o lavoro poco qualificato a tempo pieno).

Le variabili union e married mostrano effetti positivi e significativi in entrambi i modelli, ma con stime più elevate nel modello a effetti casuali, indicando che l'affiliazione sindacale e lo stato civile sono associati a salari più alti, anche al netto delle caratteristiche individuali. La differenza nei coefficienti è plausibilmente dovuta al fatto che il modello a effetti fissi controlla per tutte le caratteristiche individuali non osservate costanti nel tempo, mentre il modello a effetti casuali ne assume l'indipendenza dai regressori.

Inoltre, solo nel modello a effetti casuali è possibile stimare l'impatto di variabili invarianti nel tempo, come l'istruzione (educ), la residenza rurale (rur), e lo stato di salute (poorhlth), che risultano di interesse teorico. In particolare, l'istruzione ha un impatto positivo e altamente significativo sui salari, confermando l'importanza del capitale umano nella determinazione della retribuzione. Le interazioni tra esperienza e le variabili union e married sono significative solo nel modello RE, indicando che gli effetti dell'esperienza possono variare in base all'affiliazione sindacale o allo stato civile, fornendo spunti di analisi più ricchi.

Nel complesso, sebbene il test di Hausman suggerisca una preferenza per il modello a effetti fissi, il modello a effetti casuali risulta più adatto agli scopi dell'analisi, in quanto consente una maggiore interpretabilità, inclusione di variabili chiave e una struttura più parsimoniosa, pur mantenendo risultati robusti e coerenti.

Applichiamo adesso un modello Gamless in quanto il reddito stimato non segue una distribuzione simmetrica, ma tende a essere asimmetrico e con coda lunga destra (right-skewed). I modelli OLS o lineari misti assumono normalmente che gli errori siano normalmente distribuiti. GAMLSS ci consente invece di specificare una distribuzione alternativa, come la lognormale (LOGNO), più adatta per il reddito.

Inoltre, a differenza dei modelli OLS che assumono varianza costante (omoscedasticità), GAMLSS permette di modellare anche la varianza come funzione di variabili esplicative. Questo è particolarmente utile quando ci aspettiamo che la variabilità del reddito cambi al variare di educazione o esperienza.

Table 7: Confronto tra modello a effetti fissi e modello a effetti casuali

	Fixed Effects (within)	Random Effects (RE)
exper	0.066***	
	(0.003)	
hours	0.000***	0.000***
	(0.000)	(0.000)
union1	0.078***	0.128***
	(0.019)	(0.023)
married1	0.065***	0.192***
	(0.018)	(0.021)
(Intercept)		0.541***
		(0.106)
educ		0.089***
		(0.009)
exper_bin		0.275***
		(0.018)
rur1		−0.017
		(0.024)
poorhlth1		−0.021
		(0.048)
exper_bin × union1		−0.053+
		(0.027)
exper_bin × married1		−0.082**
		(0.025)
Num.Obs.	4360	4360
R2	0.183	0.134
R2 Adj.	0.066	0.132
RMSE	0.33	0.36

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

```r
# Carica la libreria GAMLSS
library(gamlss)

# Modello GAMLSS sul reddito stimato, con distribuzione lognormale (LOGNO)
# Stima location (mu) in funzione di educazione ed esperienza
mod_gamlss <- gamlss(
  income_estimated ~ educ + exper + hours + union + married,
  sigma.formula = ~ educ + exper + union + married,  # modelliamo anche la varianza
  data = data_rid,
  family = LOGNO()  # Distribuzione lognormale
)
```

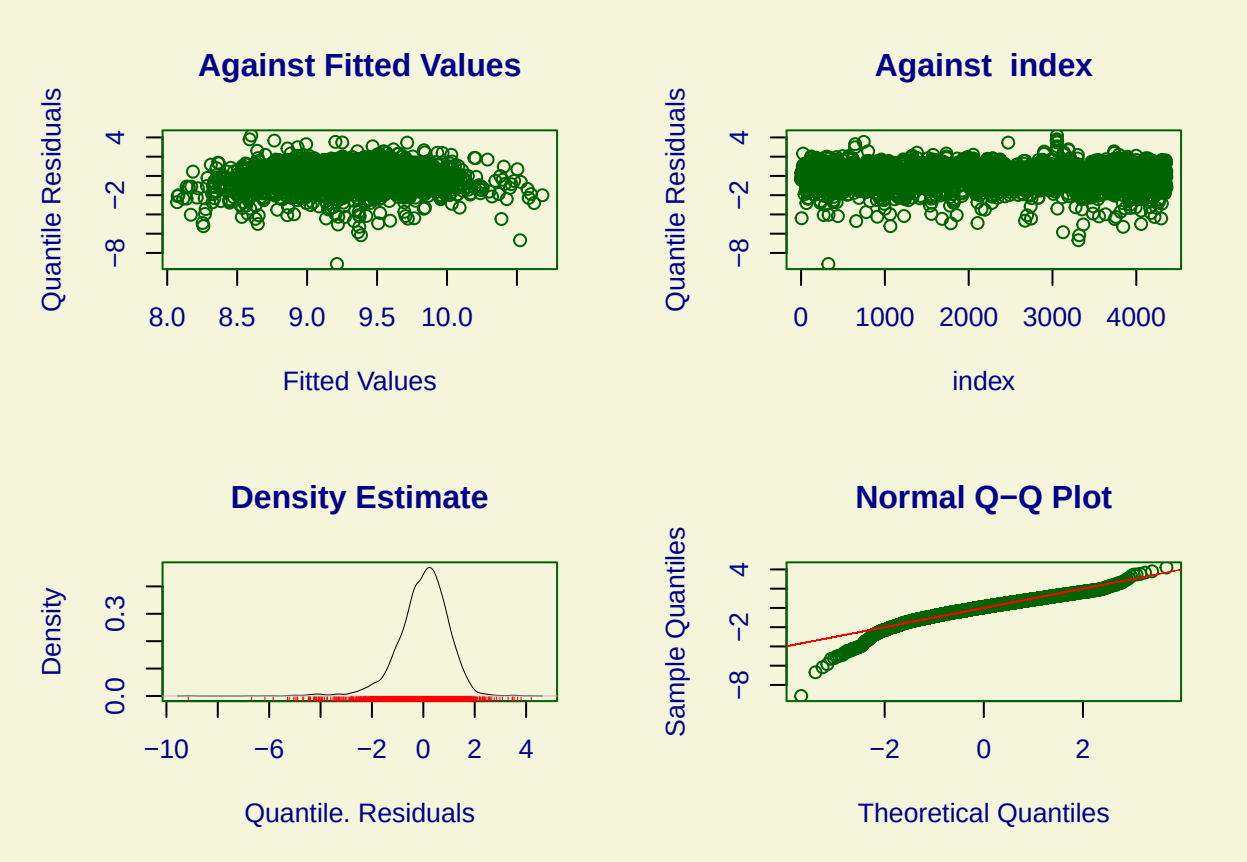
```
## GAMLSS-RS iteration 1: Global Deviance = 87392.28
## GAMLSS-RS iteration 2: Global Deviance = 87377.07
## GAMLSS-RS iteration 3: Global Deviance = 87376.91
## GAMLSS-RS iteration 4: Global Deviance = 87376.91
## GAMLSS-RS iteration 5: Global Deviance = 87376.91
```

```r
# Sommario del modello
summary(mod_gamlss)
```

```
## ******************************************************************
## Family:  c("LOGNO", "Log Normal")
##
## Call:  gamlss(formula = income_estimated ~ educ + exper +
##     hours + union + married, sigma.formula = ~educ +
##     exper + union + married, family = LOGNO(), data = data_rid)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  identity
## Mu Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.806e+00  6.774e-02 100.475   <2e-16 ***
## educ        1.014e-01  4.479e-03  22.648   <2e-16 ***
## exper       5.153e-02  2.959e-03  17.416   <2e-16 ***
## hours       3.938e-04  1.447e-05  27.209   <2e-16 ***
## union1      1.700e-01  1.625e-02  10.464   <2e-16 ***
## married1    1.447e-01  1.571e-02   9.211   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.661567   0.091709  -7.214 6.39e-13 ***
## educ         0.019905   0.006742   2.952  0.00317 **
## exper       -0.026945   0.004306  -6.257 4.31e-10 ***
## union1      -0.127864   0.025238  -5.066 4.22e-07 ***
## married1    -0.149984   0.023320  -6.431 1.40e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## ------------------------------------------------------------------
## No. of observations in the fit:  4360
## Degrees of Freedom for the fit:  11
##       Residual Deg. of Freedom:  4349
##                       at cycle:  5
##
## Global Deviance:     87376.91
##             AIC:     87398.91
##             SBC:     87469.09
## ******************************************************************
```

```
# Diagnostica grafica
plot(mod_gamlss)
```




```
## ******************************************************************
##        Summary of the Quantile Residuals
##                             mean   =  -0.001015046
##                         variance   =  1.000228
##                 coef. of skewness  =  -1.116283
##                 coef. of kurtosis  =  7.929169
## Filliben correlation coefficient   =  0.9714438
## ******************************************************************
```

Il modello GAMLSS stimato con distribuzione lognormale (LOGNO) rappresenta una scelta metodologica adatta per modellare il reddito stimato (positiva, asimmetrica). La convergenza del modello è avvenuta rapidamente entro 5 iterazioni, con una devianza globale finale pari a 87.376,91, indicativa della bontà del fit complessivo.

Dal punto di vista della media, tutte le covariate incluse nel modello, educazione, esperienza, ore lavorate, appartenenza sindacale e stato civile,risultano altamente significative ($p < 0.001$). In particolare, un anno in più di istruzione è associato a un aumento del reddito stimato di circa 0.10 unità, mentre l'esperienza contribuisce con un coefficiente positivo di circa 0.05. Anche l'affiliazione sindacale e l'essere sposati si associano positivamente al reddito, con impatti rispettivi di +0.17 e +0.14.

Un elemento distintivo del GAMLSS è la modellazione della dispersione, che in questo caso è stata specificata come funzione delle stesse variabili. I risultati evidenziano una varianza significativamente più bassa tra soggetti sindacalizzati e sposati, come indicano i coefficienti negativi su union e married nel modello per sigma. Ciò suggerisce che il reddito tra questi gruppi è non solo più alto, ma anche meno variabile, una dinamica compatibile con la maggiore stabilità occupazionale di queste categorie. Al contrario, l'istruzione è associata a una maggiore dispersione del reddito (coeff. positivo su educ), plausibilmente riflettendo le diverse opportunità economiche disponibili tra lavoratori più istruiti.

I quantile residuals mostrano una media prossima a zero e varianza pari a 1, coerente con un buon adattamento. Tuttavia, gli indicatori di skewness (-1.11) e kurtosis (7.93) rivelano una certa asimmetria e presenza di code pesanti nei residui. Il coefficiente di correlazione di Filliben (0.97) conferma una buona aderenza del modello ai dati.

L'analisi grafica dei quantile residuals del modello GAMLSS evidenzia una buona, anche se non perfetta, adeguatezza del modello. Il grafico dei residui rispetto ai valori predetti (in alto a sinistra) mostra una dispersione relativamente uniforme lungo l'asse delle ordinate, senza pattern sistematici evidenti, suggerendo che il modello non viola grossolanamente l'ipotesi di omoscedasticità. Tuttavia, si notano alcune code particolarmente estese verso il basso.

Il grafico dei residui rispetto all'indice (in alto a destra) non mostra autocorrelazione o strutture temporali particolari: i residui appaiono distribuiti in modo casuale, confermando l'indipendenza.

Il grafico di densità (in basso a sinistra) rivela una distribuzione leggermente asimmetrica e leptocurtica: i residui sono concentrati attorno allo zero ma con code più pesanti rispetto alla normale, coerentemente con i valori elevati di skewness (-1.11) e kurtosis (7.93) osservati nella diagnostica numerica.

Infine, il Q-Q plot conferma queste osservazioni, mostrando deviazioni dalla linea di riferimento sia nelle code inferiori che superiori. Questo indica che, nonostante il modello riesca a catturare bene il centro della distribuzione, fatica a modellare correttamente le code, probabilmente a causa di osservazioni estreme e dell'eterogeneità residua non spiegata.

In sintesi, il modello GAMLSS conferma l'importanza di educazione, esperienza e fattori socio-demografici nella determinazione del reddito, offrendo in più una lettura approfondita sulla variabilità individuale. Tuttavia, potrebbe beneficiare di ulteriori raffinamenti o della scelta di una distribuzione alternativa per una modellazione più precisa delle code.