ALMA MATER STUDIORUM – UNIVERSITA' DI BOLOGNA

DIPARTIMENTO DI SCIENZE STATISTICHE

"PAOLO FORTUNATI"

Corso di Laurea in Scienze Statistiche

DETERMINANTI DELLA SPERANZA DI VITA:UN'ANALISI GLOBALE DEGLI EFFETTI DI INQUINAMENTO, ABITUDINI DI CONSUMO E CONDIZIONI SANITARIE

Utilizzo Statistico di Banche Dati Online

Lucrezia Galli

2025-02-20

Presentata da:
Lucrezia Galli
Matricola: 000113963

Relatore:
Prof Paolo Verme

APPELLO I

ANNO ACCADEMICO 24 / 25

1 Introduction

Analizzare i fattori che influenzano la speranza di vita è essenziale per sviluppare strategie efficaci di sanità pubblica e promuovere il benessere a livello globale. La letteratura esistente ha messo in evidenza il ruolo di variabili socioeconomiche, dell'accesso alle cure mediche e delle abitudini di vita nella determinazione della longevità, ma una visione integrata che consideri contemporaneamente diversi fattori rimane ancora poco esplorata. Questo studio si propone di esaminare l'impatto dell'inquinamento atmosferico, del consumo di alcol e tabacco, della mortalità materna, degli incidenti stradali e delle malattie croniche sulla speranza di vita, fornendo un'analisi completa e approfondita.

2 Data

Il dataset utilizzato per l'analisi non è stato scaricato direttamente da una singola fonte, ma costruito a partire da diversi dataset disponibili nei database della World Health Organization World Health Organization (WHO) Ogni dataset conteneva una singola variabile osservata per più anni; per garantire coerenza temporale, sono stati selezionati i dati relativi all'anno 2019. Successivamente, le variabili di interesse sono state aggregate in un unico file Excel, creando così una nuova tabella strutturata. Questa è stata poi importata in R per l'analisi statistica.

Table 1: Descrizione delle variabili chiave del dataset

Variabile	Descrizione	Unita_di_misura
LifeExpatBirth	Speranza di vita alla nascita	Anni
Mortality_30_70_CCDR	Probabilità di morire tra i 30 e i 70 anni a causa di malattie croniche	Tasso (%)
PM25_Annual_Urban	Concentrazione media annua di particolato fine (PM2.5) nell'aria delle aree urbane	µg/m^3
Maternal_Mortality_Ratio	Numero di decessi materni per complicazioni legate alla gravidanza o al parto	Decessi per 100.000 nati vivi
Alcohol_Consumption	Quantità media annua di alcol consumata dalla popolazione di età superiore ai 15 anni	Litri pro capite
Tobacco_Age_Stand	Percentuale della popolazione che fa uso di prodotti a base di tabacco	Percentuale (%)

Pop_Basic_Drinking_Water	Percentuale della popolazione con accesso a servizi idrici di base	Percentuale (%)
Road_traffic_injuries	Tasso di mortalità dovuto a incidenti stradali	Tasso per 100.000 abitanti

3 Methodology

Per esaminare i fattori che influenzano l'aspettativa di vita alla nascita, è stato adottato un modello di regressione lineare multipla. Questo metodo consente di stimare l'effetto di più variabili esplicative sulla variabile dipendente, tenendo costanti gli altri fattori. Il modello è stato applicato ai dati della World Health Organization (WHO) e include variabili legate alla salute pubblica, all'ambiente e agli stili di vita. L'obiettivo è valutare l'impatto di indicatori come il tasso di mortalità prematura (Mortality_30_70_CCDR), il livello di inquinamento atmosferico urbano (PM25_Annual_Urban) e l'accesso all'acqua potabile di base (Pop_Basic_Drinking_Water) sull'aspettativa di vita nei diversi Paesi. Il modello di regressione è il seguente:

$$LifeExp = \beta_0 + \beta_1 \cdot Mort3070 + \beta_2 \cdot PM25 + \beta_3 \cdot MatMort + \beta_4 \cdot AlcCons + \beta_5 \cdot DrinkWater + \beta_6 \cdot RoadInj + \beta_7 \cdot TobUse + \epsilon$$

dove LifeExpatBirth rappresenta l'aspettativa di vita alla nascita, mentre le altre variabili sono i fattori esplicativi. Il termine indica l'errore residuo del modello. La bontà del modello è stata valutata attraverso il coefficiente di determinazione (R^2), mentre il Variance Inflation Factor (VIF) è stato calcolato per individuare eventuali problemi di multicollinearità. Inoltre, sono stati condotti test diagnostici sui residui per verificare la normalità e l'assenza di eteroschedasticità.

4 Results

```
Call:
lm(formula = LifeExpatBirth ~ Mortality_30_70_CCDR + PM25_Annual_Urban +
    Maternal_Mortality_Ratio + Alcohol_Consumption + Pop_Basic_Drinking_Water +
    Road_traffic_injuries + Tobacco_Age_Stand, data = who_data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0888 -0.8150  0.2523  1.1095  4.7449

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)              76.364472   2.089773  36.542  < 2e-16 ***
Mortality_30_70_CCDR     -0.511643   0.025220 -20.287  < 2e-16 ***
PM25_Annual_Urban        -0.026790   0.012352  -2.169  0.03160 *
Maternal_Mortality_Ratio -0.007184   0.001477  -4.865 2.76e-06 ***
Alcohol_Consumption      -0.030798   0.047600  -0.647  0.51856
Pop_Basic_Drinking_Water  0.104331   0.017929   5.819 3.23e-08 ***
Road_traffic_injuries    -0.123732   0.021235  -5.827 3.10e-08 ***
Tobacco_Age_Stand         0.054323   0.018374   2.957  0.00359 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.09 on 157 degrees of freedom
  (16 osservazioni eliminate a causa di valori mancanti)
Multiple R-squared:  0.9198,    Adjusted R-squared:  0.9162
F-statistic: 257.2 on 7 and 157 DF,  p-value: < 2.2e-16
```

```
vif(lmod)
```

```
    Mortality_30_70_CCDR        PM25_Annual_Urban Maternal_Mortality_Ratio
                1.469175                 1.344254                 2.967496
     Alcohol_Consumption Pop_Basic_Drinking_Water    Road_traffic_injuries
                1.368504                 3.120186                 1.994744
       Tobacco_Age_Stand
                1.242602
```

Il modello di regressione lineare multipla utilizzato ha spiegato il 91,98% della variazione nell'aspettativa di vita alla nascita (LifeExpatBirth), come indicato dal valore di R^2 (0.9198). Il test F ha restituito un valore di 257.2 con un p-value inferiore a 2.2e-16, confermando la significatività globale del modello. Questi risultati suggeriscono che le variabili indipendenti selezionate hanno un forte potere esplicativo nel determinare l'aspettativa di vita.

```
Tabella 1: Coefficienti della regressione
================================================
                            Dependent variable:
                        ------------------------
                              LifeExpatBirth
------------------------------------------------
Mortality_30_70_CCDR              -0.512***
```

```
                                  (0.025)

PM25_Annual_Urban                -0.027**
                                  (0.012)

Maternal_Mortality_Ratio         -0.007***
                                  (0.001)

Alcohol_Consumption               -0.031
                                  (0.048)

Pop_Basic_Drinking_Water         0.104***
                                  (0.018)

Road_traffic_injuries            -0.124***
                                  (0.021)

Tobacco_Age_Stand                0.054***
                                  (0.018)

Constant                         76.364***
                                  (2.090)

-------------------------------------------------
Observations                        165
R2                                 0.920
Adjusted R2                        0.916
Residual Std. Error          2.090 (df = 157)
F Statistic               257.151*** (df = 7; 157)
=================================================
Note:                     *p<0.1; **p<0.05; ***p<0.01
```

4.1 Interpretazione dei Coefficienti

Dai coefficienti della regressione, si osserva che l'aumento del tasso di mortalità tra i 30 e i 70 anni (Mortality_30_70_CCDR) è associato a una riduzione significativa dell'aspettativa di vita (-0.512 per unità di incremento, $p<0.001$). Anche la qualità dell'aria (PM25_Annual_Urban) incide negativamente, seppur con un effetto più contenuto (-0.027, $p<0.05$). Variabili come il tasso di mortalità materna (Maternal_Mortality_Ratio) e gli incidenti stradali (Road_traffic_injuries) hanno un impatto negativo significativo. Al contrario, un miglior accesso all'acqua potabile

(Pop_Basic_Drinking_Water) è positivamente correlato con l'aspettativa di vita. L'uso di tabacco (Tobacco_Age_Stand) mostra un effetto positivo, sebbene meno marcato.

4.2 Multicollinearità e Robustezza del Modello

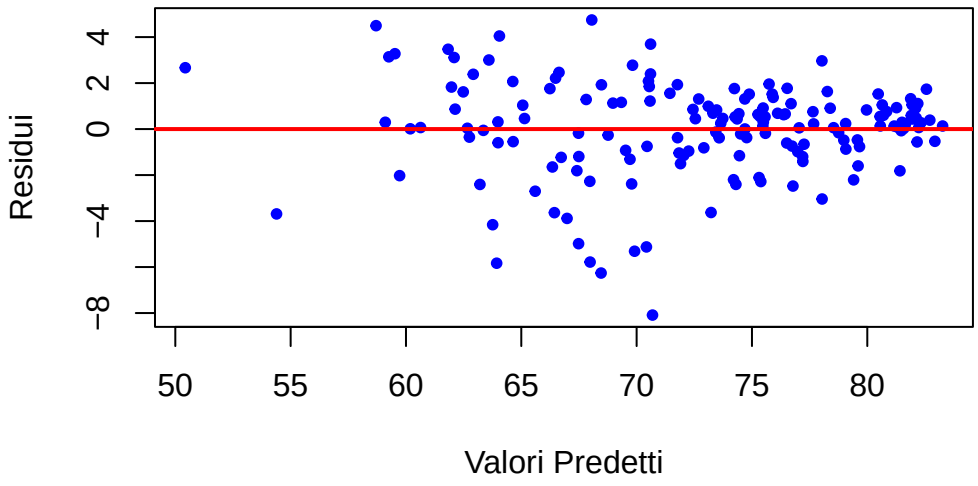
Per verificare la robustezza del modello, sono stati calcolati i valori VIF per ogni variabile indipendente. Nessuna variabile ha un VIF superiore a 5, indicando l'assenza di gravi problemi di multicollinearità.

```
vif_table <- data.frame(Variable = names(vif(lmod)), VIF = vif(lmod))
```

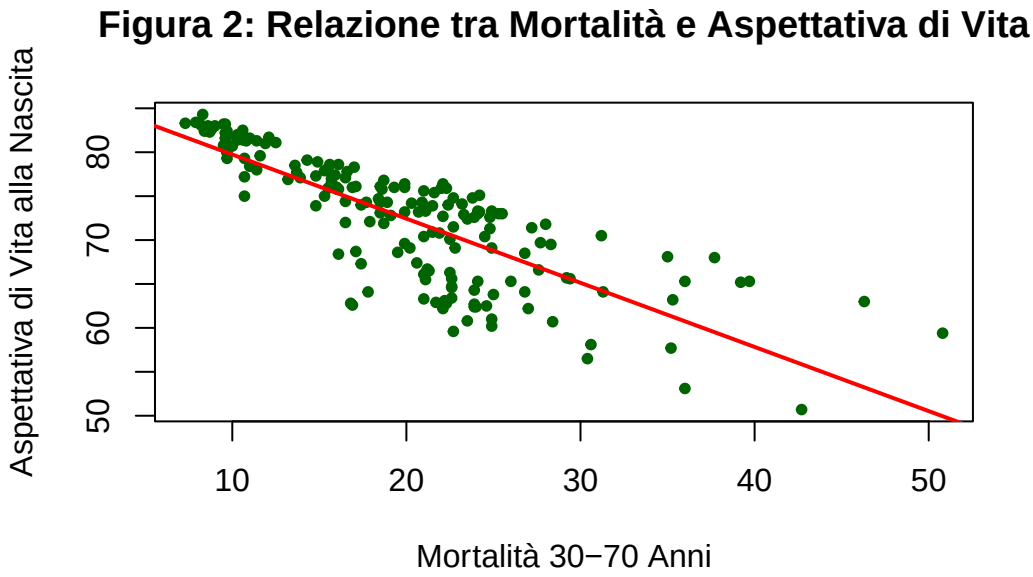
4.3 Visualizzazione dei risultati

```
plot(lmod$fitted.values, resid(lmod),
     xlab = "Valori Predetti",
     ylab = "Residui",
     main = "Figura 1: Residui vs. Valori Predetti",
     pch = 20, col = "blue")
abline(h = 0, col = "red", lwd = 2)
```

Figura 1: Residui vs. Valori Predetti



```r
plot(who_data$Mortality_30_70_CCDR, who_data$LifeExpatBirth,
     xlab = "Mortalità 30-70 Anni",
     ylab = "Aspettativa di Vita alla Nascita",
     main = "Figura 2: Relazione tra Mortalità e Aspettativa di Vita",
     pch = 20, col = "darkgreen")
abline(lm(LifeExpatBirth ~ Mortality_30_70_CCDR, data = who_data), col = "red", lwd =
```



This study is based on statistical analysis methods presented by Smith (smith2020?) and recent developments in data science discussed by Jones and Brown (jones2018?).

5 Machine Learning in Healthcare

Lee and White (lee2019?) highlighted key advancements in AI-driven healthcare, while Williams (williams2021?) examined societal impacts of big data.

6 Online Resources

For a general understanding of statistical learning, see Doe's online guide (doe2023?).