Un'analisi statistica della Sindrome dell'Ovaio Policistico: individuazione di fattori di rischio e sviluppo di un modello diagnostico

Francesca Bellissimo

Matricola: 247295

Contents

1 Introduzione

La sindrome dell'ovaio policistico (PCOS, dall'inglese *Polycystic Ovary Syndrome*) è un disturbo endocrino-metabolico molto diffuso tra le donne in età fertile; la prevalenza complessiva risulta essere del 6% (95% IC: 5%-8%) secondo i criteri diagnostici del *National Institute of Health* (NIH), del 10% (95% IC: 8%-13%) secondo i criteri della *Androgen Excess & PCOS Society*, e del 10% (95% IC: 7%-13%) secondo i criteri di Rotterdam [1].

È importante precisare che, a dispetto del nome, la presenza di cisti ovariche non è un requisito indispensabile per la diagnosi di PCOS. In molti casi, le donne con PCOS presentano numerosi piccoli follicoli che non completano il processo di maturazione, piuttosto che cisti vere e proprie. La PCOS, comunque, è associata a numerosi sintomi, eterogenei e non sempre tutti presenti. Uno dei tratti maggiormente distintivi è il verificarsi di disfunzioni ovariche, manifestate tramite la presenza di oligomennorea ($35 <$ durata del ciclo mestruale in giorni < 90) o amenorrea (durata del ciclo mestruale in giorni > 90). Tali condizioni derivano, in genere, da anovulazione cronica e determinano spesso infertilità. Va specificato, tuttavia, che il 60% delle donne con PCOS sono fertili, sebbene il tempo necessario al concepimento risulti maggiore [2]. Le donne con PCOS presentano frequentemente un rapporto LH:FSH (ormone luteinizzante/ormone follicolo-stimolante) elevato, una condizione che contribuisce all'aumento della produzione di androgeni, responsabili a loro volta di sintomi come irsutismo, acne e alopecia androgenetica. L'insulino-resistenza è un altro tratto molto comune nelle donne con PCOS; in presenza di tale condizione, le cellule risultano scarsamente sensibili all'insulina e, di conseguenza, il pancreas ne produce quantità maggiori. Questo porta a iperinsulinemia, che a sua volta può contribuire all'accumulo di tessuto adiposo e impedisce il completo sviluppo dei follicoli ovarici. Inoltre, la *International Diabetes Federation* ha identificato la PCOS come un fattore di rischio significativo e non modificabile associato al Diabete di tipo 2 [3]. Oltre agli aspetti biologici, la sindrome è associata anche a problemi psicologici, come disturbi d'ansia e depressione, che si verificano con maggiore frequenza nelle donne con PCOS, soprattutto tra quelle non fertili, rispetto alla popolazione generale [4].

Il trattamento della PCOS si concentra principalmente sulla gestione dei sintomi e sulla correzione dei disturbi ormonali e metabolici. In prima linea, si consiglia una sostanziale modifica dello stile di vita. È possibile migliorare la sensibilità all'insulina e ripristinare un normale funzionamento mestruale tramite l'esercizio fisico e una modesta perdita di peso [5]. Il trattamento farmacologico, invece, prevede principalmente l'utilizzo di contraccettivi orali o di sensibilizzanti all'insulina (come la Metformina), sebbene tali metodi non debbano sostituire l'adozione di uno stile di vita più sano.

2 Materiali e obiettivi dell'analisi

Gli obiettivi di questa analisi sono i seguenti:

- determinare l'impatto del sovrappeso e dell'obesità sul rischio di sviluppare la PCOS;
- analizzare più in profondità il ruolo del peso corporeo, valutando se il rischio di PCOS sia legato specificamente al sovrappeso e all'obesità o, più in generale, a un peso non equilibrato (incluso l'essere sottopeso);
- indagare la relazione tra il rapporto LH:FSH e il BMI;
- sviluppare un classificatore per la PCOS in grado di distinguere tra i casi di anovulazione e quelli di ovulazione regolare;
- valutare la relazione tra il numero di follicoli e il rapporto LH:FSH tramite un modello di regressione di Poisson.

Ho utilizzato il *Polycystic Ovary Sindrome Dataset* [6], disponibile su Kaggle. Il dataset include informazioni relative a 541 pazienti di dieci diversi ospedali dello stato del Kerala, in India. Tra queste, 177 (il 33% circa) presentano la sindrome. Il dataset contiene 42 variabili, sia cliniche che relative allo stile di vita; una breve descrizione e relative statistiche descrittive sono riportate nella tabella seguente.

Table 1: Statistiche descrittive delle variabili

	Tipo	Descrizione	mean	sd	min	max
PCOS	binaria	Presenza di PCOS (0 = No, 1 = Sì)	0.33	0.47	0.00	1.00
Age	discreta	Età(anni)	31.43	5.41	20.00	48.00
Weight	continua	Peso(kg)	59.64	11.03	31.00	108.00
Height	continua	Altezza(cm)	156.48	6.03	137.00	180.00
BMI	continua	Indice di massa corporea(kg/mˆ2)	24.32	4.05	12.42	38.90
Blood.Group	discreta	Gruppo sanguigno	13.80	1.84	11.00	18.00
Pulse.rate	continua	Frequenza cardiaca(bpm)	73.46	2.69	70.00	82.00
RR	continua	Intervallo RR(respiri/min)	19.24	1.69	16.00	28.00
Hb	continua	Emoglobina(g/dl)	11.16	0.87	8.50	14.80
Cycle	binaria	Ciclo mestruale irregolare (0 = No, 1 = Sì)	0.28	0.45	0.00	1.00
Cycle.length	discreta	Lunghezza della mestruazione(giorni)	4.94	1.49	0.00	12.00
Marriage.status	continua	Anni di matrimonio	7.68	4.80	0.00	30.00
Pregnant	binaria	Essere incinta (0 = No, 1 = Sì)	0.38	0.49	0.00	1.00
Aborptions	discreta	Numero di aborti	0.29	0.69	0.00	5.00
I.beta.HCG	continua	Prima betaHCG(mIU/mL)	664.55	3348.92	1.30	32460.97
II.beta.HCG	continua	Seconda betaHCG(mIU/mL)	238.23	1603.83	0.99	25000.00
FSH	continua	Ormone follicolo-stimolante(mIU/mL)	14.60	217.02	0.21	5052.00
LH	continua	Ormone luteinizzante(mIU/mL)	6.47	86.67	0.02	2018.00
LHFSH	continua	Rapporto FSH/LH	0.55	0.45	0.00	4.35
Hip	continua	Circonferenza dei fianchi(pollici)	37.99	3.97	26.00	48.00
Waist	continua	Circonferenza della vita(pollici)	33.84	3.60	24.00	47.00
WHR	continua	Rapporto vita/fianchi	0.89	0.05	0.76	0.98
TSH	continua	Ormone tireostimolante(mIU/mL)	2.98	3.76	0.04	65.00

	Tipo	Descrizione	mean	sd	min	max
AMH	continua	Ormone anti-mulleriano(ng/mL)	5.62	5.88	0.10	66.00
PRL	continua	Prolattina(ng/mL)	24.32	14.97	0.40	128.24
Vit.D3	continua	Vitamina D3(ng/mL)	49.92	346.21	0.00	6014.66
PRG	continua	Progesterone(ng/mL)	0.61	3.81	0.05	85.00
RBS	continua	Glicemia (mg/dl)	99.84	18.56	60.00	350.00
Weight.gain	binaria	Aumento di peso (0 = No, 1 = Sì)	0.38	0.49	0.00	1.00
Hair.growth	binaria	Irsutismo (0 = No, 1 = Sì)	0.27	0.45	0.00	1.00
Skin.darkening	binaria	Iperpigmentazione (0 = No, 1 = Sì)	0.31	0.46	0.00	1.00
Hair.loss	binaria	Alopecia (0 = No, 1 = Sì)	0.45	0.50	0.00	1.00
Pimples	binaria	Acne (0 = No, 1 = Sì)	0.49	0.50	0.00	1.00
Fast.food	binaria	Dieta non salutare (0 = No, 1 = Sì)	0.52	0.50	0.00	1.00
Reg.Exercise	binaria	Esercizio regolare (0 = No, 1 = Sì)	0.25	0.43	0.00	1.00
BP.systolic	continua	Pressione sistolica(mmHg)	114.66	7.38	12.00	140.00
BP.diastolic	continua	Pressione diastolica (mmHg)	76.93	5.57	8.00	100.00
FolliclesL	discreta	Numero di follicoli nell'ovaio sinistro	6.13	4.23	0.00	22.00
FolliclesR	discreta	Numero di follicoli nell'ovaio destro	6.64	4.44	0.00	20.00
L.size	continua	Dimensione media dei follicoli nell'ovaio sinistro	15.02	3.57	0.00	24.00
R.size	continua	Dimensione media dei follicoli nell'ovaio destro	15.45	3.32	0.00	24.00
Endometrium	continua	Dimensione dell'endometrio(mm)	8.48	2.17	0.00	18.00

Nel corso dell'analisi, ho aggiunto alcune variabili derivate da quelle già presenti nel dataset. Ne riporto una breve descrizione di seguito.

- Follicles: è una variabile discreta che rappresenta il massimo tra il numero di follicoli nell'ovaia sinistra e nell'ovaia destra.
- BMIlevel: è una variabile categoriale di quattro categorie (*Underweight*, *Normal Weight*, *Overweight*, *Obese*), costruite rispettando le soglie segnalate dall'OMS [7].
- HighBMI: è una variabile binaria che vale 1 se la paziente è sovrappeso o obesa, cioè ha un indice di massa corporea maggiore o uguale di 25, 0 altrimenti.
- HighFHLSH: è una variabile binaria che vale 1 se la paziente ha un rapporto FH:LSH elevevato (> 2), 0 altrimenti.
- PCOStype: è una variabile categoriale di tre categorie (*None*, *Anovulatory*, *Ovulatory*), definite secondo il livello dell'ormone AMH.

Sono stati rilevati alcuni valori mancanti. Per garantire la completezza dei dati, questi sono stati sostituiti con la mediana della variabile corrispondente, una scelta basata sulla robustezza della mediana in presenza di valori estremi.

3 Analisi e risultati

3.1 Relazione tra la PCOS e il peso corporeo

Per valutare il legame fra avere la PCOS ed essere sovrappeso o obese, ho inserito le variabili PCOS e HighBMI in una tabella di contingenza.

```
##         PCOS
## HighBMI   1   0 Sum
##     0    83 231 314
##     1    94 133 227
##     Sum 177 364 541
```

Di seguito, una tabella che evidenzia le probabilità stimate relative a ciascuna combinazione.

```
##         PCOS
## HighBMI         1         0
##       0 0.2643312 0.7356688
##       1 0.4140969 0.5859031
```

Dunque, la probabilità stimata di avere la PCOS se si è sottopeso o normopeso è

$$\hat{\pi}_0 \approx 0.26.$$

Di seguito, il relativo intervallo di confidenza asintotico al 95%.

```
binom.confint(x = table[1,1], n = 314, conf.level = 0.95, methods = c("asymptotic"))
```

```
##       method  x   n      mean     lower     upper
## 1 asymptotic 83 314 0.2643312 0.2155561 0.3131064
```

Invece, la probabilità stimata di avere la PCOS se si ha un BMI elevato è

$$\hat{\pi}_1 \approx 0.41.$$

Di seguito, il relativo intervallo di confidenza asintotico al 95%.

```
binom.confint(x = table[2,1], n = 227, conf.level = 0.95, methods = c("asymptotic"))
```

```
##       method  x   n      mean     lower     upper
## 1 asymptotic 94 227 0.4140969 0.3500204 0.4781735
```

La differenza tra le due proporzioni è

$$\hat{\pi}_0 - \hat{\pi}_1 \approx -0.15, \quad 95\% \text{ Wald CI: } (-0.23, -0.07).$$

Si noti che l'intervallo di confindenza asintotico per la differenza non comprende lo 0, pertanto le due probabilità risultano significativamente diverse. Tale risultato è confermato anche dal test del χ^2 per le seguenti ipotesi:

$$\begin{aligned} H_0: &\quad \pi_0 - \pi_1 = 0 \\ H_a: &\quad \pi_0 - \pi_1 \neq 0. \end{aligned}$$

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  table[, 1] out of rowSums(table)
## X-squared = 13.425, df = 1, p-value = 0.0002483
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.23029408 -0.06923733
## sample estimates:
##    prop 1    prop 2
## 0.2643312 0.4140969
```

Una stima del rischio relativo

$$RR = \frac{\hat{\pi}_1}{\hat{\pi}_0} \approx 1.57, \quad 95\% \text{ Wald CI: } (1.23, 1.99).$$

È possibile affermare, allora, che la probabilità stimata di avere la PCOS per le donne sovrappeso o obese è del 57% maggiore rispetto alle donne sottopeso o normopeso. In genereale, con una fiducia del 95%, è possibile affermare che per le donne sovrappeso o obese la probabilità di avere la PCOS è maggiore di una percentuale che va dal 23% al 99% rispetto alle donne sottopeso o normopeso. Si noti, inoltre, che l'intervallo di confidenza per il rischio relativo non contiene 1, pertanto le due probabilità risultano significativamente diverse.

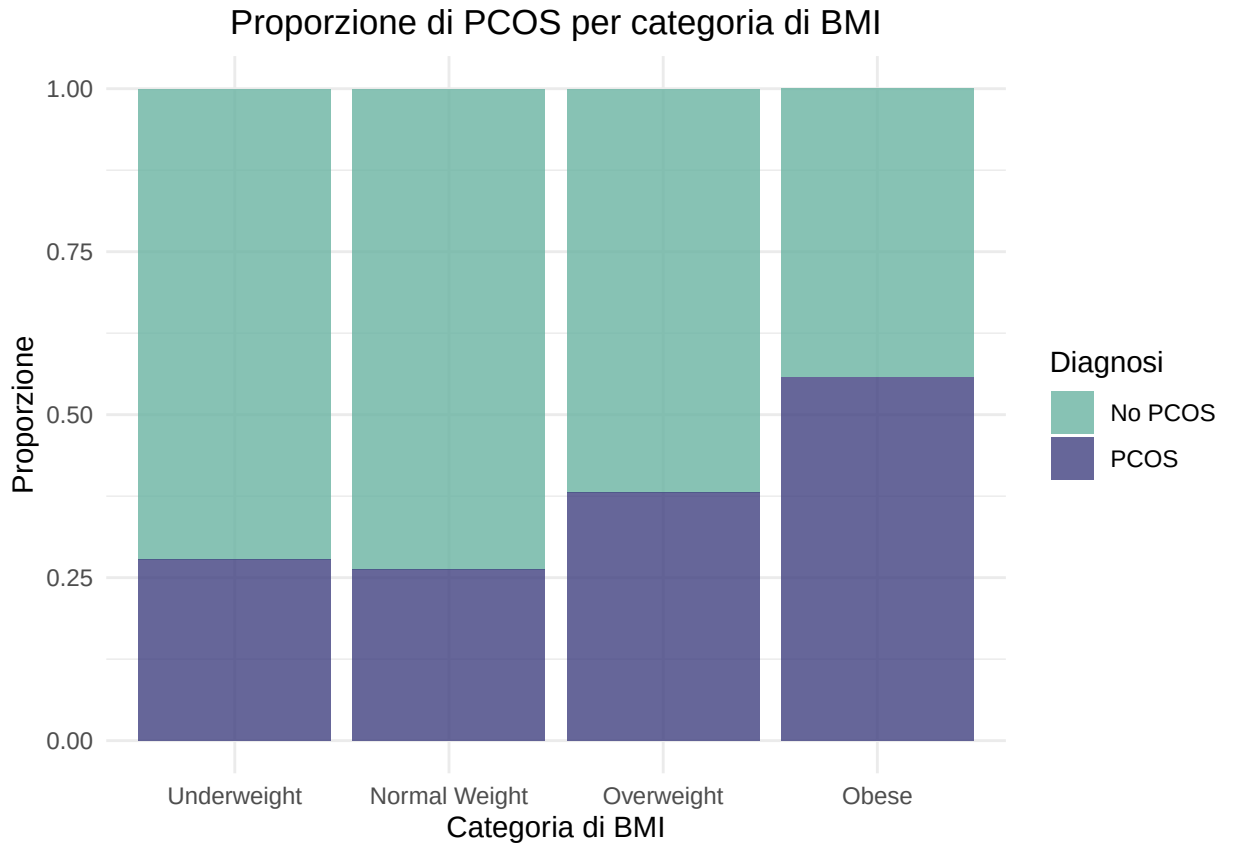
Una stima dell'*odds ratio*, invece, è

$$OR = \frac{odds_1}{odds_0} \approx 1.97, \quad 95\% \text{ Wald CI: } (1.37, 2.83).$$

L'*odds* di avere la PCOS per donne sovrappeso o obese è 1.97 volte l'*odds* di avere la PCOS per donne sottopeso o normopeso. Si osservi che l'intervallo di confidenza per l'*odds ratio* non include 1, pertanto i due *odds* risultano significativamente diversi.

Dall'analisi di queste misure è possibile concludere che la distribuzione della PCOS nei due gruppi è significativamente diversa; la PCOS sembra essere più comune nelle donne sovrappeso o obese che in quelle sottopeso o normopeso.

Per esplorare più a fondo questa relazione, è possibile valutare la distribuzione della PCOS per ogni categoria di BMI. Il grafico seguente mostra la proporzione di donne con PCOS per ogni livello della variabile *BMIlevel*. Emerge una relazione complessivamente crescente tra *BMIlevel* e la probabilità di avere la PCOS. Le proporzioni di PCOS sono leggermente più alte per la categoria *Underweight* rispetto a *Normal Weight*, ma mostrano un incremento marcato nelle categorie *Overweight* e *Obese*.



Successivamente, ho stimato un modello di regressione logistica utilizzando la variabile *BMIlevel* come esplicativa, intendendo *Normal Weight* come categoria di riferimento.

```
##
## Call:
## glm(formula = PCOS ~ BMIlevel, family = binomial(link = logit),
##     data = pcos)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.03255    0.13630  -7.576 3.57e-14 ***
## BMIlevelUnderweight  0.07704    0.39628   0.194 0.845858
## BMIlevelOverweight   0.54485    0.20404   2.670 0.007580 **
## BMIlevelObese        1.26617    0.33597   3.769 0.000164 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 683.99  on 540  degrees of freedom
## Residual deviance: 666.14  on 537  degrees of freedom
## AIC: 674.14
##
## Number of Fisher Scoring iterations: 4
```

I risultati sono interessanti. Come previsto, le categorie *Overweight* e *Obese* sono significativamente associate a un aumento del rischio di PCOS rispetto alla categoria di riferimento *NormalWeight*. In particolare, *Obese* risulta essere la categoria con l'impatto più marcato. Per la categoria *Underweight*, invece, non si evidenzia alcuna relazione significativa rispetto alla categoria di riferimento; non emergono evidenze sufficienti per concludere che essere sottopeso influenzi il rischio di PCOS. I risultati sono concordi al contesto clinico: sebbene il sottopeso possa essere associato a disfunzioni di vario genere, la PCOS è più comunemente correlata a condizioni come l'insulino-resistenza, che tendono ad avere una maggiore prevalenza nei soggetti con sovrappeso o obesità.

Di seguito sono riportati gli *odds ratio* e i relativi intervalli di confidenza. Essere sovrappeso o obese, piuttosto che essere normopeso, aumenta l'*odds* di avere la PCOS dell'72% e del 255%, rispettivamente. Sebbene entrambe le condizioni aumentino il rischio di sviluppare la sindrome rispetto al normopeso, i risultati mostrano che l'obesità ha un impatto aggiuntivo significativo rispetto al sovrappeso. Infatti, confrontando direttamente queste due categorie, si osserva che l'*odds* di avere la PCOS nelle donne obese rispetto a quelle sovrappeso è più che raddoppiato. Per quanto riguarda, invece, la categoria *Underweight*, l'*odds ratio* non risulta significativamente diverso da 1.

```
##  contrast                    odds.ratio    SE  df asymp.LCL asymp.UCL
##  Underweight / Normal Weight       1.08 0.428 Inf     0.497      2.35
##  Overweight / Normal Weight        1.72 0.352 Inf     1.156      2.57
##  Overweight / Underweight          1.60 0.642 Inf     0.726      3.51
##  Obese / Normal Weight             3.55 1.190 Inf     1.836      6.85
##  Obese / Underweight               3.28 1.580 Inf     1.276      8.45
##  Obese / Overweight                2.06 0.705 Inf     1.051      4.03
##
## Confidence level used: 0.95
## Intervals are back-transformed from the log odds ratio scale
```

3.2 Relazione tra LH:FSH e BMI

Tra le strategie comunemente raccomandate per la gestione della PCOS, la riduzione del BMI gioca un ruolo cruciale, come si è dimostrato nella precedente sezione. Tuttavia, non tutti i sintomi possono essere ricondotti esclusivamente a un peso corporeo elevato. In particolare, il rapporto LH:FSH tra l'ormone luteinizzante e

quello follicolo-stimolante richiede dei trattamenti diversi. Per dimostrare ciò, ho costruito una tabella di contingenza 2×4 inserendo la variabile *HighFHLSH* e la variabile *BMIlevel* e includendo soltanto i dati relativi alle pazienti con PCOS.

```
##          BMIlevel
## HighFHLSH Normal Weight Underweight Overweight Obese Sum
##       0              71          10         69    22 172
##       1               2           0          1     2   5
##       Sum            73          10         70    24 177
```

Di seguito, la tabella che riporta le proporzioni relative a ogni categoria.

```
##          BMIlevel
## HighFHLSH Normal Weight Underweight  Overweight       Obese
##         0   0.401129944 0.056497175 0.389830508 0.124293785
##         1   0.011299435 0.000000000 0.005649718 0.011299435
```

Ho testato l'indipendenza tra le due variabili tramite il test del χ^2 e il test del rapporto tra le verosimiglianze. Entrambi sono concordi nel supportare l'ipotesi di indipendenza. Questo risultato indica che le variazioni del rapporto ormonale non sono direttamente influenzate dal valore del BMI.

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 2.9398  3  0.40100
## Pearson          3.4427  3  0.32827
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.138
## Cramer's V        : 0.139
```

Dunque, sebbene un indice di massa corporea elevato sia un fattore di rischio significativo per la PCOS e possa influenzare diversi sintomi, i dati analizzati evidenziano che un rapporto LH:FSH elevato deve essere gestito con altre strategie.

Tuttavia, bisogna segnalare che quest'ultima analisi potrebbe non essere completamente robusta. In particolare, l'approssimazione al χ^2 potrebbe non essere pienamente valida a causa della presenza di numerosità inferiori a 5:

```
ind.test$expected
```

```
##          BMIlevel
## HighFHLSH Normal Weight Underweight Overweight      Obese
##         0     70.937853   9.7175141  68.022599 23.3220339
##         1      2.062147   0.2824859   1.977401  0.6779661
```

3.3 Classificatore per il tipo di PCOS

L'obiettivo di questa sezione è costruire un modello di regressione multinomiale che sia in grado di prevedere accuratamente la classe di *PCOStype* alla quale la paziente appartiene.

Per definire la PCOS con anovulazione, ho considerato i livelli dell'ormone anti-Mulleriano (AMH) come criterio diagnostico, stabilendo che valori superiori a 5 ng/ml indicano anovulazione, mentre valori inferiori a questa soglia suggeriscono il contrario [8] [9].

Alcune variabili sono state eliminate dal dataset per evitare ridondanze o per la presenza di correlazioni elevate con altre variabili. In particolare, sono state rimosse le seguenti:

- *Weight* e *Height*, mantenendo il sintentico indicatore *BMI*;
- *Marriage Status*, poiché altamente correlata ad *Age*;
- *FSH* e *LH*, mantenendo il più informativo rapporto *FSH/LH*, che fornisce una misura più adeguata dell'equilibro ormonale;
- *Waist* e *Hip*, in favore del più rappresentativo rapporto WHR (*waist to hip ratio*), che sintetizza meglio la distribuzione del grasso corporeo;
- *L.size* e *R.size*, poiché direttamente legate a *F.size*, che rappresenta la minima dimensione tra i follicoli dell'ovaio sinistro e dell'ovaio destro;
- *FolliclesL* e *FolliclesR* poiché direttamente legate a *Follicles*, che rappresenta il numero massimo di follicoli tra le due ovaie;
- *PCOS* e *AMH* poiché utilizzate per la definizione della variabile target *PCOStype*.

Per l'addestramento del modello, ho suddiviso il dataset in un set di addestramento (80%) e un set di test (20%), facendo in modo che le proporzioni dei diversi tipi di PCOS fossero mantenute in modo equilibrato in entrambi i sottoinsiemi. Di seguito sono riportate le proporzioni dei diversi tipi di PCOS nel training set e nel test set. Si noti che le classi presentano un evidente sbilanciamento, il che potrebbe influenzare le prestazioni del modello.

```
## Proporzioni nel training set:
```

```
##
##        None Anovulatory   Ovulatory
##   0.6728111   0.1866359   0.1405530
```

```
##
## Proporzioni nel test set:
```

```
##
##        None Anovulatory   Ovulatory
##   0.6728972   0.1869159   0.1401869
```

Il dataset contiene, a questo punto, 31 possibili variabili esplicative. Per la loro selezione, ho utilizzato la tecnica della forward selection utilizzando il BIC come criterio di scelta.

Il modello finale avrà la seguente forma:

$$\log(\pi_j/\pi_{\text{None}}) = \beta_{j0} + \beta_{j1}X_1 + \cdots + \beta_{jp}X_p, \text{ per } j = \text{Anovulatory, Ovulatory,}$$

e le variabili $X_1, \ldots, X_p$ sono quelle selezionate dalla procedura descritta sopra.

Le variabili selezionate dalla forward selection sono: *Follicles, Weight.Gain, Hair.growth, Cycle, Skin.darkening.*

```
## Call:
## multinom(formula = formula(forw.sel2), data = train_data, trace = FALSE)
##
## Coefficients:
##             (Intercept) Follicles Weight.gain Hair.growth    Cycle
## Anovulatory   -6.971301 0.4601188    1.538351    1.349762 1.656117
## Ovulatory     -7.877094 0.5160472    1.604352    1.528190 1.168259
##             Skin.darkening
## Anovulatory       1.201199
## Ovulatory         1.341118
##
## Residual Deviance: 408.3781
## AIC: 432.3781
```

Di seguito, i risultati di un test di valutazione globale per i parametri stimati. Tutte le variabili risultano significative.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: PCOStype
##                LR Chisq Df Pr(>Chisq)
## Follicles       143.962  2  < 2.2e-16 ***
## Weight.gain      19.650  2  5.409e-05 ***
## Hair.growth      13.708  2  0.0010552 **
## Cycle            17.363  2  0.0001697 ***
## Skin.darkening   11.862  2  0.0026557 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nelle tabelle seguenti, sono riportate delle metriche di valutazione delle performance del modello sul test set.

Table 2: Metriche per classe

Metric	None	Anovulatory	Ovulatory
Sensitivity	0.9722	0.6500	0.3333
Specificity	0.8571	0.8966	0.9456
Precision	0.9333	0.5909	0.5000
F1-score	0.9524	0.6190	0.4000

Table 3: Metriche globali

Metric	Value
Accuracy	0.8224
F1-score Micro	0.8224

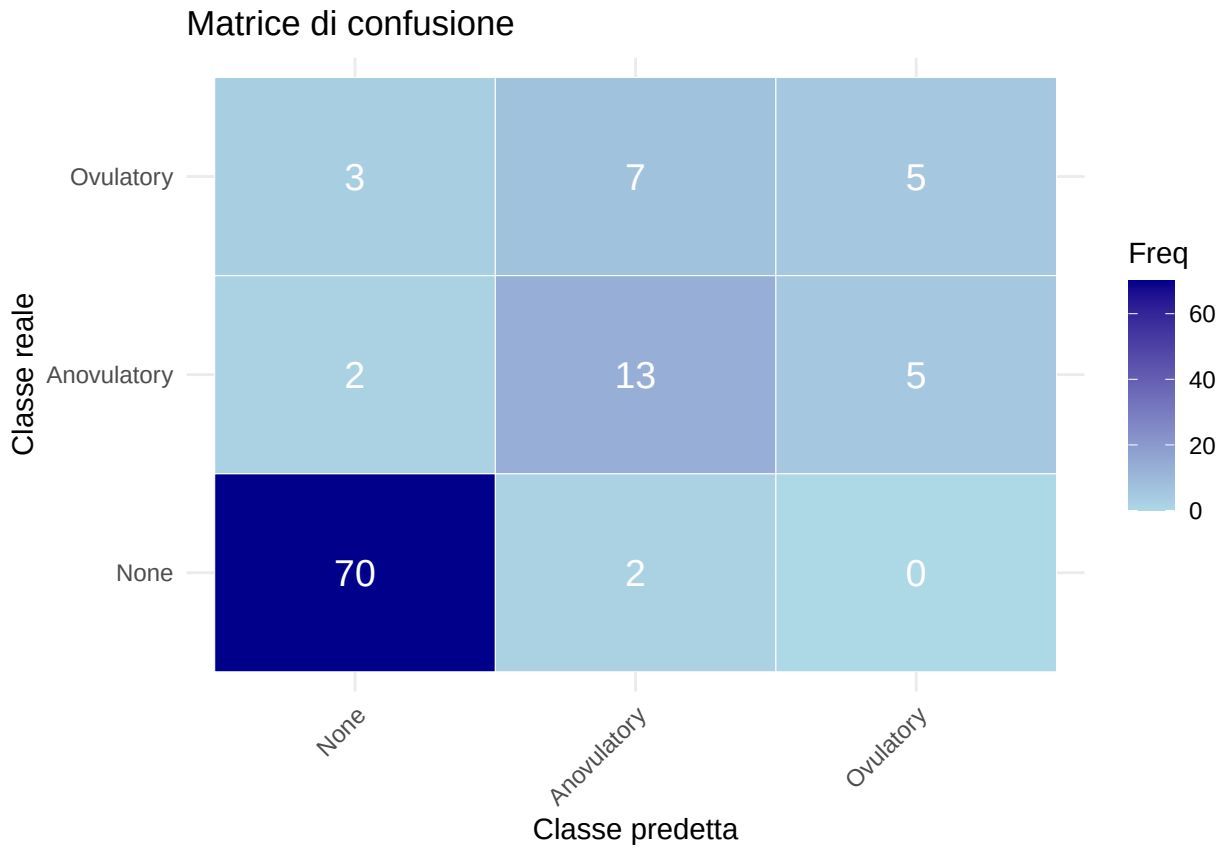
Emergono alcune connessioni tra le diverse classi e le prestazioni del modello. La classe *None* si distingue per le sue ottime performance, con una *sensitivity* di 0.9722 e una *specificity* di 0.8571. L'F1-score di 0.9524 conferma che per questa classe il bilanciamento tra *precision* e *sensitivity* è eccellente, suggerendo che il modello è molto affidabile nel riconoscere questa categoria. Questo risultato potrebbe essere dovuto sia alla chiara separabilità della classe None rispetto alle altre due, sia alla sua maggiore rappresentatività nel dataset, che facilita l'apprendimento del modello.

D'altra parte, le performance per le classi *Anovulatory* e *Ovulatory* sono significativamente inferiori. Per la classe *Anovulatory*, la *sensitivity* è pari a 0.65, indicando che il modello fatica a riconoscere correttamente tutti i casi di questa categoria. Tuttavia, la *specificity* di 0.8966 suggerisce che il modello è altamente preciso nell'escludere i casi che non appartengono alla classe *Anovulatory.* La *precision* per questa classe è relativamente bassa (0.5909), suggerendo che molte delle predizioni per *Anovulatory* sono errate. Il risultato complessivo, sintetizzato da una F1-score di 0.6190, indica che la capacità del modello di distinguere questa classe è migliorabile.

La classe *Ovulatory* è quella più problematica. La *sensitivity* di appena 0.3333 dimostra che il modello è in grado di identificare solo un terzo dei veri casi di *Ovulatory*, mancando molti esempi reali di questa condizione. Sebbene la *specificity* sia elevata (0.9457), il che indica che il modello evita di assegnare erroneamente questa classe quando non dovrebbe, la *precision* di 0.5 e l'F1-score di 0.4 mostrano che le predizioni per *Ovulatory* sono altamente inaffidabili. Questo suggerisce che il modello ha difficoltà a distinguere questa classe da *Anovulatory*, probabilmente a causa di una sovrapposizione nelle caratteristiche delle due categorie o di una minore rappresentatività di *Ovulatory* nel dataset di addestramento.

Nel complesso, l'*accuracy* del modello è buona (0.8224), ma le differenze nelle performance tra le classi indicano un problema di sbilanciamento. La classe *None* è nettamente più facile da classificare, mentre il modello mostra difficoltà nell'identificare correttamente *Anovulatory* e, soprattutto, *Ovulatory.* Il valore di F1-score micro (0.8224) conferma che il modello è efficace nel complesso, ma non equamente performante tra le classi. Questo valore riflette la media ponderata degli F1-score considerando la dimensione di ciascuna classe, quindi è influenzato dal fatto che la classe None è preponderante.

Di seguito, è riportata la matrice di confusione.



3.4 Modello di regressione di Poisson per il numero di follicoli

L'obiettivo di questa sezione è sviluppare un modello di regressione di Poisson per il numero di follicoli nelle donne con PCOS, utilizzando il rapporto LH:FSH come variabile esplicativa. L'intento non è fornire un predittore accurato del numero di follicoli né di spiegare completamente le variazioni osservate. Piuttosto, si vuole esaminare il modo in cui il numero di follicoli varia in funzione del rapporto ormonale, contribuendo a una migliore comprensione della sindrome.

La variabile *Follicles* è discreta e positiva ma, prima di procedere, è necessario verificare che la sua distribuzione sia sufficientemente vicina a una distribuzione di Poisson, per la quale si verifica che media e varianza coincidono. Tuttavia, per la variabile *Follicles* si ha

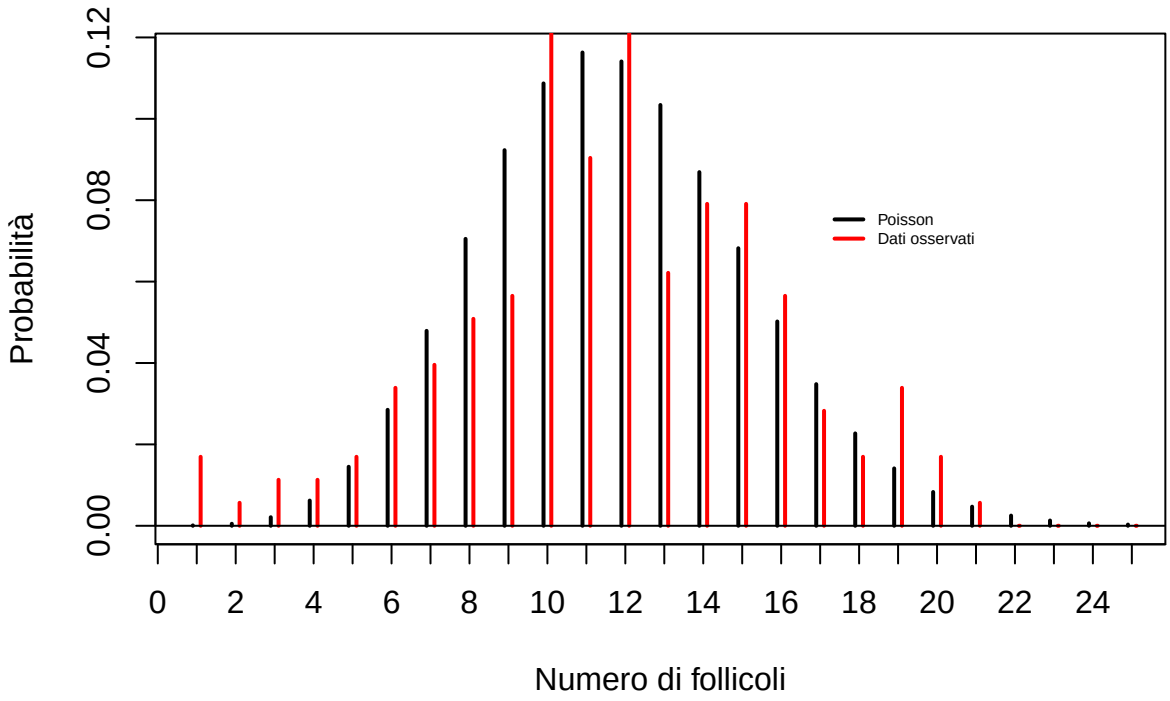
```
mean(pcos_subset$Follicles)
```

```
## [1] 11.77401
```

```
var(pcos_subset$Follicles)
```

```
## [1] 17.22137
```

La differenza tra la media e la varianza di *Follicles* suggerisce che la variabile non segue esattamente una distribuzione di Poisson. Il grafico seguente, che riporta il confronto tra la distribuzione empirica di *Follicles* e quella teorica di una variabile di Poisson (con media pari a quella di *Follicles*), conferma questo risultato.



Procediamo comunque alla stima del modello, ma bisognerà tenere in conto che questo disallineamento ne intaccherà la robustezza.

Il modello che si vuole stimare è

$$\mu = e^{\beta_0 + \beta_1 LHFSH},$$

dove μ è la media di *Follicles*. Il risultato del processo di stima è riportato di seguito.

```
##
## Call:
## glm(formula = Follicles ~ LHFSH, family = poisson(link = log),
##     data = pcos_subset)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.57566    0.03555  72.444  < 2e-16 ***
## LHFSH       -0.17718    0.04673  -3.791  0.00015 ***
## ---
```

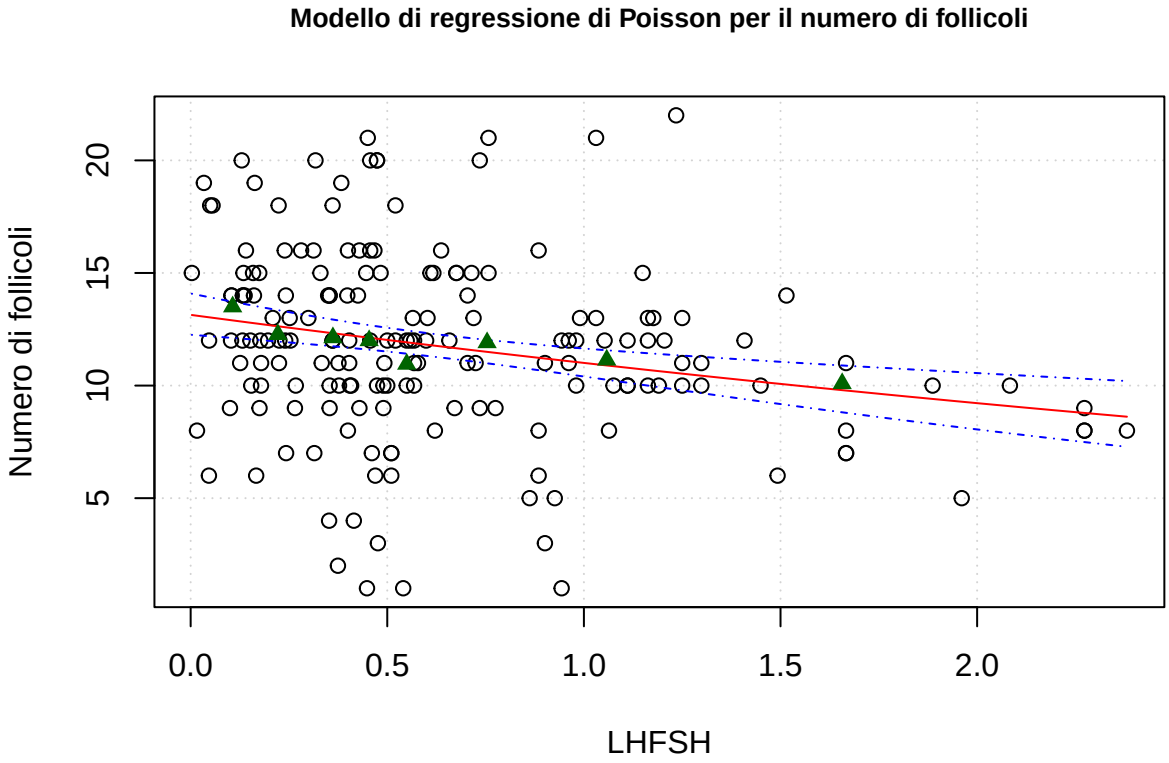
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 283.54  on 176  degrees of freedom
## Residual deviance: 268.57  on 175  degrees of freedom
## AIC: 1021.9
##
## Number of Fisher Scoring iterations: 4
```

In corrispondenza di un aumento di 0.5 unità (la deviazione standard) del rapporto LH:FSH, il numero medio atteso di follicoli varia di un valore percentuale pari a $100(e^{\beta_1} - 1)\% \approx -8.46\%$. Di seguito, l'intervallo di confidenza al 95% di tale variazione.

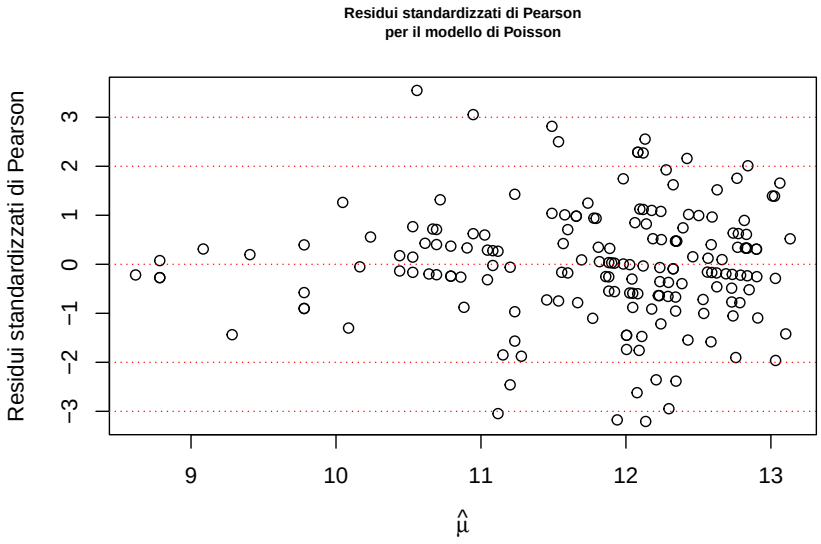
```
## Waiting for profiling to be done...
```

```
##      2.5 %     97.5 %
## -12.596448  -4.226959
```

La curva riportata nel seguente grafico rappresenta il modello di regressione stimato, evidenziando l'andamento del numero medio di follicoli al variare del rapporto LH:FSH. Per ottenere una visualizzazione più chiara, i valori di LH:FSH sono stati suddivisi in otto intervalli, ognuno contenente il 12.5% delle osservazioni. All'interno di ciascun intervallo, i valori medi del numero di follicoli sono stati calcolati e rappresentati dai triangoli verdi nel grafico.



Il grafico mostra che la relazione tra LH:FSH e il numero medio di follicoli è negativa. Il modello sembra complessivamente adeguato, ma alcune medie si trovano al di fuori dell'intervallo di confidenza. La valutazione dei residui standardizzati di Pearson, infatti, evidenzia alcuni problemi. Numerosi residui presentano valori in modulo superiori a 2 (o anche a 3 in alcuni casi), il che suggerisce la presenza di osservazioni che non sono ben rappresentate dal modello.



Una delle soluzioni al problema della sovradispersione è l'utilizzo di un modello di regressione quasi-Poisson. L'idea è quella di prevedere che media e varianza non siano uguali, ma legate dalla relazione

$$Var(Y) = \gamma E(Y),$$

dove γ è una costante che rappresenta la dispersione. Il processo di stima dei parametri conduce a stime identiche a quelle ottenute con il modello di Poisson standard, ma l'inclusione del parametro γ (che deve anch'esso essere stimato) consente di tenere conto esplicitamente del fenomeno della sovradispersione, correggendo così le misure diagnostiche.

La stima del parametro di dispersione γ è riportata nell'output e risulta essere

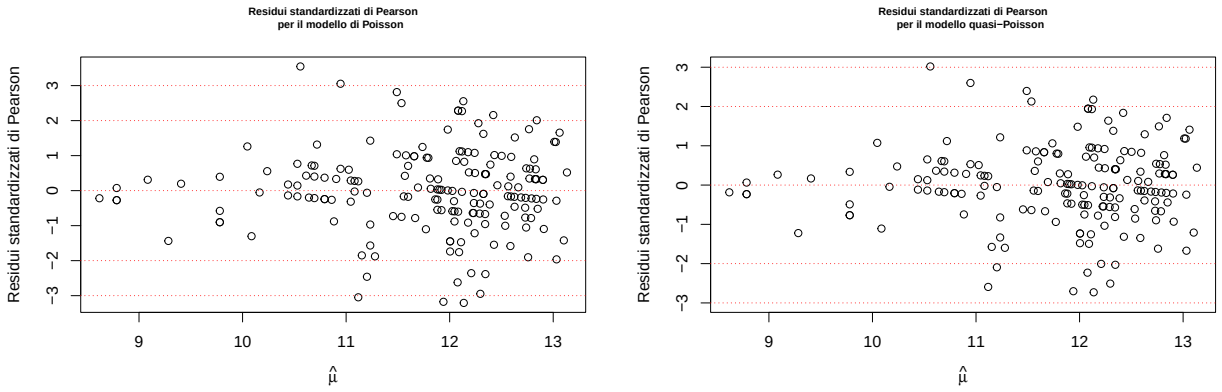
$$\hat{\gamma} \approx 1.38.$$

Si noti che, per un modello quasi-Poisson, non è possibile calcolare l'AIC, che infatti non è riportato. Sebbene sia possibile calcolare il QAIC, questa misura non sarebbe appropriata per confrontare il modello quasi-Poisson con il modello di regressione di Poisson precedentemente stimato. Per tale motivo, il calcolo del QAIC non è stato incluso.

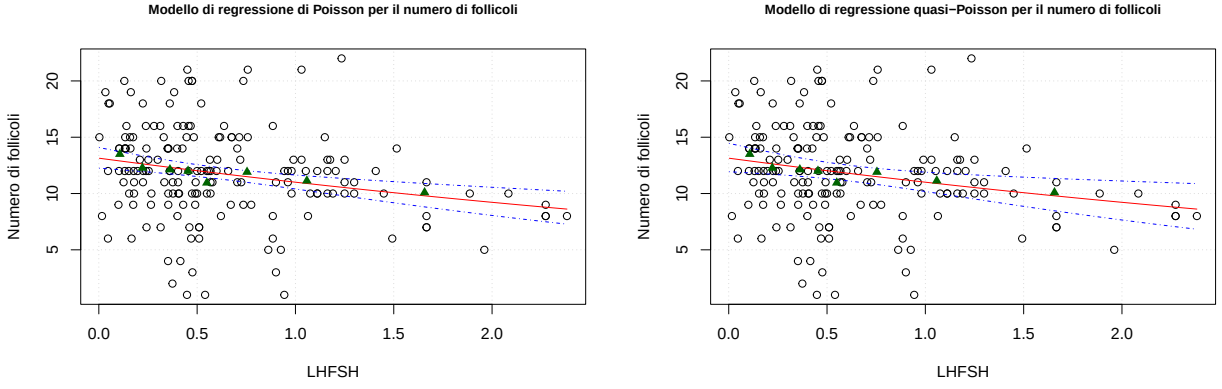
```
##
## Call:
## glm(formula = Follicles ~ LHFSH, family = quasipoisson(link = log),
##     data = pcos_subset)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.57566    0.04177  61.659  < 2e-16 ***
## LHFSH       -0.17718    0.05490  -3.227  0.00149 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.380404)
##
##     Null deviance: 283.54  on 176  degrees of freedom
## Residual deviance: 268.57  on 175  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Di seguito è riportato il confronto fra i residui standardizzati di Pearson per il modello di regressione di Poisson e quello quasi-Poisson. Il numero di residui maggiori di 2 in valore assoluto è nettamente diminuito; inoltre, è presente un solo residuo maggiore di 3.



Dai grafici seguenti è possibile osservare che l'introduzione del parametro di dispersione non modifica le stime dei parametri, ma si riflette in una più accurata stima della loro varianza. Per tale ragione, l'intervallo di confidenza nel caso di un modello quasi-Poisson risulta più ampio.



I risultati di questa analisi mostrano una relazione negativa tra il rapporto LH:FSH e il numero di follicoli nelle donne con PCOS. Questo risultato potrebbe riflettere specificità del campione oppure fattori biologici

sottostanti. Ad esempio, un rapporto LH:FSH elevato potrebbe indicare la presenza di disfunzioni più severe, a causa delle quali le ovaie non rispondono in modo normale agli stimoli ormonali, riducendo il numero di follicoli osservabili. Va sottolineato, tuttavia, che i livelli di FH e LSH sono soggetti a significative fluttuazioni durante il ciclo mestruale; non è noto se le misurazioni riportate nel dataset sono state effettuate nella stessa fase mestruale per tutte le pazienti. Inoltre, bisogna tenere conto del fatto che altre variabili (come l'età, l'indice di massa corporea, l'AMH) hanno una certa influenza sulla relazione tra LH:FSH e numero di follicoli. Pertanto, per una comprensione più precisa, sono necessarie analisi più approfondite basate su dati standardizzati in termini di fase mestruale e che tengano conto di eventuali variabili confondenti.

4 Conclusioni

Questo studio ha esaminato diversi aspetti della Sindrome dell'Ovaio Policistico. Per quanto riguarda il legame tra PCOS e peso corporeo, i risultati confermano che l'obesità è un fattore di rischio significativo per lo sviluppo della sindrome. Questi risultati supportano l'importanza della gestione del peso come parte del trattamento e della prevenzione della PCOS.

Tuttavia, l'analisi del rapporto LH:FSH ha rivelato che l'indice di massa corporea, pur essendo un fattore rilevante per altri aspetti della sindrome, non ha un impatto diretto sul rapporto ormonale. Il test di indipendenza tra BMI e LH:FSH non ha mostrato relazioni significative, suggerendo che il trattamento delle alterazioni ormonali potrebbe richiedere approcci specifici e diversi dalla semplice gestione del peso.

L'analisi ha evidenziato una relazione negativa tra il rapporto LH:FSH e il numero di follicoli nelle donne con PCOS, suggerendo che un rapporto elevato potrebbe indicare disfunzioni ovariche più gravi. Tuttavia, l'interpretazione è limitata dalla variabilità nei livelli di LH e FSH durante il ciclo mestruale e dall'assenza di dati standardizzati in termini di fase mestruale. Studi futuri basati su dati più accurati e che considerino altre variabili confondenti, come età e AMH, sono necessari per approfondire questa relazione e migliorarne la comprensione.

La costruzione di un modello di regressione multinomiale per la previsione del tipo di PCOS ha prodotto risultati modesti. Con un'accuratezza dell'82.24%, il modello ha mostrato una moderata capacità di classificazione complessiva. Tuttavia, l'analisi delle metriche per classe ha evidenziato un forte squilibrio nelle prestazioni. Il modello, certamente, potrebbe beneficiare di miglioramenti per gestire meglio le classi meno frequenti.

Inoltre, è necessario tenere in considerazione alcune limitazioni. In primo luogo, non sono disponibili informazioni precise riguardo alla strategia di campionamento e questo potrebbe ridurre l'applicabilità dei risultati a popolazioni diverse. Inoltre, non è stato possibile indagare la presenza di eventuali fattori di confondimento, come l'uso di farmaci o la presenza di altre patologie, e questo potrebbe comportare dei bias nei risultati.

Sarebbe interessante, inoltre, disporre di dati longitudinali, che consentirebbero di esaminare come le relazioni tra le variabili e la PCOS evolvono nel tempo. Soprattutto, sarebbe possibile osservare l'effetto di trattamenti o modifiche dello stile di vita sullo stato della sindrome, permettendo di valutare se e come questi interventi possano migliorare i sintomi e prevenire complicazioni future.

5 Bibliografia

1. Bozdag G, Mumusoglu S, Zengin D, Karabulut E, Yildiz BO. The prevalence and phenotypic features of polycystic ovary syndrome: A systematic review and meta-analysis. Human reproduction. 2016;31:2841–55.

2. Brassard M, AinMelk Y, Baillargeon J-P. Basic infertility including polycystic ovary syndrome. Medical Clinics of North America. 2008;92:1163–92.

3. Alberti KGMM, Zimmet P, Shaw J. International diabetes federation: A consensus on type 2 diabetes prevention. Diabetic Medicine. 2007;24:451–63.

4. Deeks AA, Gibson-Helm ME, Teede HJ. Anxiety and depression in polycystic ovary syndrome: A comprehensive investigation. Fertility and sterility. 2010;93:2421–3.

5. Huber-Buchholz M-M, Carey D, Norman R. Restoration of reproductive potential by lifestyle modification in obese polycystic ovary syndrome: Role of insulin sensitivity and luteinizing hormone. The Journal of Clinical Endocrinology & Metabolism. 1999;84:1470–4.

6. Kottarathil P. Polycystic ovary syndrome (PCOS). 2020.

7. WHO Consultation on Obesity (1999: Geneva S, Organization WH. Obesity : Preventing and managing the global epidemic : Report of a WHO consultation. 2000;252 p.

8. Dewailly D, Gronier H, Poncelet E, Robin G, Leroy M, Pigny P, et al. Diagnosis of polycystic ovary syndrome (PCOS): Revisiting the threshold values of follicle count on ultrasound and of the serum AMH level for the definition of polycystic ovaries. Human Reproduction. 2011;26:3123–9.

9. Bhattacharya K, Saha I, Sen D, Bose C, Chaudhuri GR, Dutta S, et al. Role of anti-mullerian hormone in polycystic ovary syndrome. Middle East Fertility Society Journal. 2022;27:32.