Esercitazione 5 - Statistica (parte II)

Davide Passaretti

9/3/2017

Contents

1 Inferenza sulla regressione semplice

Nell'ambito della statistica inferenziale, il modello di regressione semplice ha il fine di stimare la **vera** retta della popolazione:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \varepsilon$$

tramite la retta determinata sul campione usando il metodo dei minimi quadrati (tale metodo è comunemente chiamato **OLS**, acronimo di **O**rdinary **L**east **S**quares):

$$\mathbf{y} = b_0 + b_1 \mathbf{x} + e$$

Se sono valide le ipotesi del teorema di Gauss-Markov formulate su ε, allora b_0 e b_1 sono i più efficienti stimatori lineari non distorti (per questo motivo chiamati **BLUE**, acronimo di **B**est **L**inear **U**nbiased **E**stimators) rispettivamente dell'intercetta e della pendenza della retta della popolazione.

Sappiamo già dalla prima parte del corso (Statistica descrittiva) che la regressione studia la variazione in media di una variabile di risposta (o dipendente) **y** come funzione lineare di una variabile esplicativa (o indipendente) **x**. Se tale variazione è significativa, vuol dire che davvero la variabile indipendente ha un potere esplicativo lineare su quella di risposta. In tal caso, la retta di regressione non sarà piatta: la media (condizionata) della **y**, cioè $\mathrm{E}(\mathbf{y}|\mathbf{x})$, crescerà (se $\beta_1 > 0$) o descrescerà (se $\beta_1 < 0$) al crescere della **x**.

1.1 Test sulla pendenza della retta

Se la retta di regressione della popolazione è piatta, non vi è alcuna relazione *lineare* tra **y** ed **x** (ma, **attenzione!** Potrebbe esistere una relazione non lineare tra le due variabili). Ciò implica che $\beta_1 = 0$. Dobbiamo quindi testare l'ipotesi nulla che la pendenza sia zero contro l'ipotesi alternativa che vi sia una relazione lineare tra **y** ed **x** (vale a dire $H_1 : \beta_1 \neq 0$).

È possibile dimostrare che la statistica test calcolata sul campione:

$$T_n = \frac{b_1 - \overbrace{\beta_1}^{=0 \text{ sotto } H_0}}{\sqrt{\frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^2}}}$$

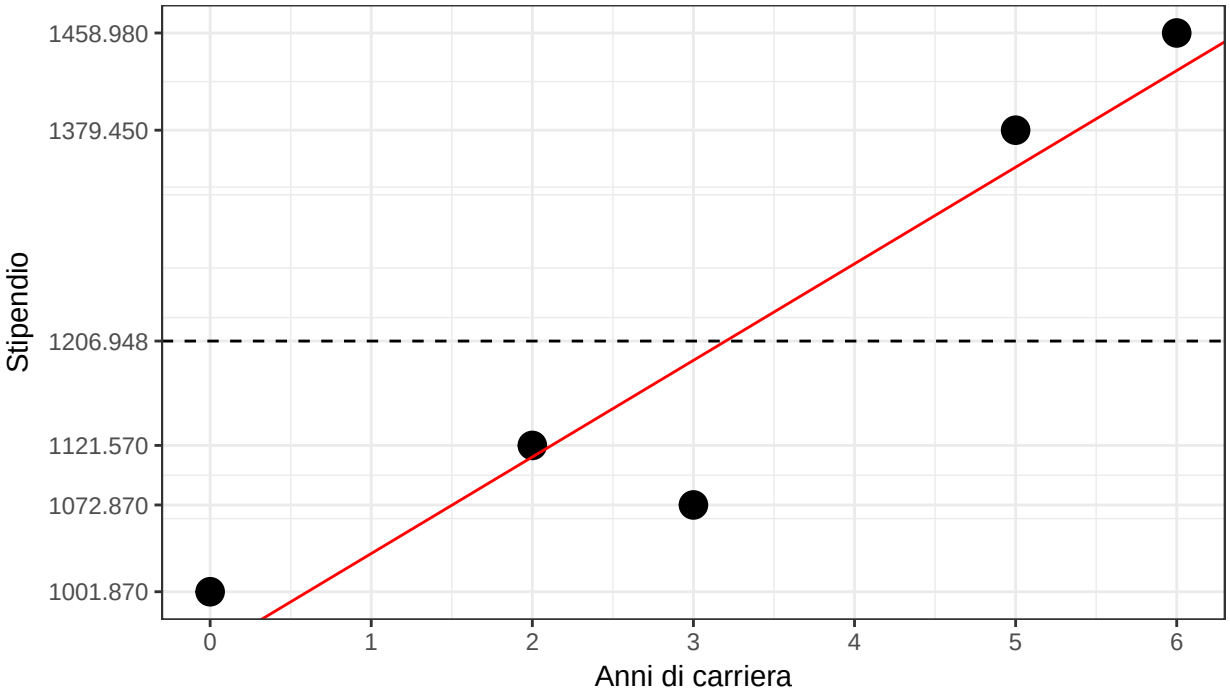
si distribuisce come una Normale standardizzata qualora le ipotesi aggiuntive di Normalità e di indipendenza degli errori siano valide. Siccome non conosciamo la varianza dell'errore, la stimiamo dai residui di regressione:

$$s_\epsilon^2 = \frac{\underbrace{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}_{\text{Dev}(e)}}{n-2}$$

Ne consegue che T_n diviene una t di Student con $n-2$ gradi di libertà.

Facciamo un esercizio:

Di seguito sono riportati in forma grafica gli anni di carriera di 5 persone collocate nel settore terziario (sulle ascisse) ed i relativi stipendi attuali mensili (sulle ordinate):



Si vuole verificare se lo stipendio in tale settore dipende linearmente dagli anni di carriera usando un livello di significatività del 5%.

Dal campione otteniamo:

- $\bar{\mathbf{x}} = 3.2$
- $\bar{\mathbf{y}} = 1206.948$
- $s_{\mathbf{xy}} = \frac{\sum(\mathbf{x}_i - 3.2)(\mathbf{y}_i - 1206.948)}{5-1} = \frac{1801.712}{4} = 450.428$
- $s_{\mathbf{x}}^2 = \frac{\text{Dev}(\mathbf{x})}{n-1} = \frac{\sum(\mathbf{x}_i - 3.2)^2}{5-1} = \frac{22.8}{4} = 5.7$

- $s^2_{\mathbf{y}} = \frac{\text{Dev}(\mathbf{y})}{n-1} = \frac{\sum(\mathbf{x}_i - 1206.948)^2}{5-1} = \frac{160600.4}{4} = 40150.09$
- $b_1 = \frac{s^2_{\mathbf{xy}}}{s^2_{\mathbf{x}}} = \frac{450.428}{5.7} = 79.022$
- $b_0 = \bar{\mathbf{y}} - b_1\bar{\mathbf{x}} = 954.076$
- $s^2_{\varepsilon} = \frac{\text{Dev}(e)}{n-2} = \frac{\sum(\mathbf{y}_i - 954.076 - 79.022\,\bar{\mathbf{x}})^2}{5-2} = \frac{18224.66}{3} = 6074.887$

L'area di non rifiuto di H_0 è:

$$\mathcal{A}_b = [t_{0.025,\,5-2},\, t_{0.975,\,5-2}] = [-3.18,\, 3.18]$$

Il valore osservato della statistica test è:

$$T_5 = \frac{79.022}{\sqrt{\frac{6074.887}{22.8}}} = 4.841 \notin \mathcal{A}_b \longrightarrow \text{rifiuto } H_0$$

1.2 Test sull'R^2

L'R^2 indica la percentuale della variabilità di **y** spiegata dal modello. Se tale percentuale nella popolazione è zero, vuol dire che la variabile indipendente non contribuisce a spiegare linearmente la variabile di risposta. Dunque, l'ipotesi nulla è $H_0 : R^2 = 0$, mentre l'ipotesi alternativa è $H_1 : R^2 > 0$. Il test è necessariamente unidirezionale: vogliamo inferire se il modello spiega una percentuale di variabilità della **y** maggiore di zero.

La statistica test (o meglio, uno dei modi per scriverla) è:

$$T_n = \frac{\hat{R}^2}{\left(\frac{1-\hat{R}^2}{n-2}\right)} \sim \mathbf{F}_{1,\,n-2}$$

Vogliamo verificare, usando $\alpha = 0.05$, l'utilità del modello di regressione stimato al punto precedente tramite un test sull'R^2.

Prima bisogna stimare l'R^2. Possiamo avvalerci di qualche informazione già calcolata sul campione:

$$\hat{R}^2 = b_1^2\frac{s^2_{\mathbf{x}}}{s^2_{\mathbf{y}}} = 79.022\frac{5.7}{40150.09} = 0.8865$$

L'area di non rifiuto di H_0 è:

$$\mathcal{A}_d = [0,\, F_{0.05,\,1,\,5-2}] = [0,\, 10.128]$$

Il valore osservato della statistica test è:

$$T_5 = \frac{0.8865}{\left(\frac{1-0.8865}{5-2}\right)} = 23.4368 \notin \mathcal{A}_d \longrightarrow \text{rifiuto } H_0$$

1.3 Equivalenza dei due test nell'ambito della regressione semplice

Partiamo dall'R^2 e ragioniamo a livello della popolazione. Tale indice è il rapporto tra la devianza di regressione (cioè la devianza spiegata dal modello) e quella totale di **y** (ovvero la somma di quella di regressione e quella dell'errore). Un R^2 nullo indica che non vi è devianza spiegata, cioè che la devianza totale coincide con la devianza dell'errore. Ciò accade quando la retta di regressione è piatta, vale a dire quando

$\beta_1 = 0$. Dunque $H_0 : R^2 = 0$ e $H_0 : \beta_1 = 0$ sono due ipotesi equivalenti nell'ambito della regressione semplice, a cui seguono test altrettanto equivalenti. A sostegno di ciò, vale la seguente relazione tra distribuzioni:

$$F_{1,\,\nu,\,\alpha} = t^2_{\nu,\,\frac{\alpha}{2}}$$

La verifica che i valori critici ed osservati degli Esercizi 1.1 ed 1.2 seguano la suddetta relazione è lasciata come esercizio.

2 Cenni di regressione lineare multipla

La regressione lineare multipla prende in considerazione più di una variabile esplicativa. Geometricamente, passiamo da un problema in 2 dimensioni (diagramma cartesiano di $\mathbf{y}$ contro $\mathbf{x}$) ad un problema in più dimensioni. Per esempio, una regressione con due variabili esplicative può raffigurarsi in uno spazio 3D, in cui bisogna trovare il piano (non più la retta!) che meglio interpoli i punti.

Aggiungiamo al nostro esempio anche l'età del lavoratore. Avremo il seguente modello di regressione stimato:

$$\textsc{Stipendio} = b_0 + b_1 \textsc{Anni di carriera} + b_2 \textsc{Età} + e$$

Di conseguenza, bisognerà stimare tre coefficienti tramite **OLS**. Le equazioni di stima sono più complesse di quelle della regressione semplice, perciò ci si avvale di un software. Per esempio, supponiamo che la persona appena entrata nel mondo del lavoro abbia 26 anni, quella con 2 anni di carriera abbia 25 anni, quella con 3 anni di carriera abbia 30 anni, quella con 5 anni di carriera abbia 31 anni e quella con 6 anni di carriera abbia 29 anni.

L'output del software `R` è il seguente:

```
##
## Call:
## lm(formula = Stipendio ~ AnniCarriera + Eta)
##
## Residuals:
##       1       2       3       4       5
##  53.316 -32.457 -82.053  56.087   5.106
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1429.98     598.11   2.391    0.139
## AnniCarriera    93.48      25.07   3.728    0.065 .
## Eta            -18.52      23.12  -0.801    0.507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.07 on 2 degrees of freedom
## Multiple R-squared:  0.9141, Adjusted R-squared:  0.8281
## F-statistic: 10.64 on 2 and 2 DF,  p-value: 0.08593
```

Possiamo eseguire principalmente 4 test.

- Un test t sulla significatività dell'intercetta β_0 (ma non ci interessa molto, come non ci interessava nel caso della regressione semplice).
  $H_0 : \beta_0 = 0$.
- Un test t sulla significatività della pendenza β_1 (il coefficiente associato agli anni di carriera).
  $H_0 : \beta_1 = 0$.

- Un test t sulla significatività della pendenza β_2 (il coefficiente associato all'età del lavoratore). $H_0 : \beta_2 = 0$.
- Un test **F** sull'R^2 per verificare la significatività congiunta delle due pendenze (bontà del modello). $H_0 : \beta_1 = \beta_2 = 0$.

2.1 Test di tipo *t* sui singoli coefficienti

Per i test di tipo t, abbiamo la colonna del `t value`, cioè la colonna contenente il valore osservato della statistica test per ciascun coefficiente. Ciascun valore osservato è il risultato del rapporto tra la stima del coefficiente (contenuta nella prima colonna – `Estimate`) e lo standard error (contenuto nella seconda colonna).

I gradi di libertà di ciascuna t non sono più $n-2$, bensì $n-3$, poiché si stanno stimando 3 coefficienti. In generale, i g.d.l. (in inglese `DF`, acronimo di *degrees of freedom*) sono $n-p$, dove p è il numero di coefficienti da stimare.

Verificare la significatività di ciascun coefficiente usando $\alpha = 10\%$.

L'area di non rifiuto di H_0 è dunque:

$$\mathcal{A}_b = [-t_{0.05,\,5-3},\, t_{0.05,\,5-3}] = [-2.92,\, 2.92]$$

Per un livello di significatività del 10 %, l'area di accettazione include b_0 e b_2, ma non b_1, quindi gli anni di carriera influenzano significativamente lo stipendio. Ciò è in linea (ovviamente) con quanto indicato dai p-value (ultima colonna – `Pr(>|t|)`).

2.2 Test F sull'R^2

L'R^2 della regressione multipla (nell'output del software è indicato con `Multiple R-squared`) è ancora la percentuale della varianza di **y** spiegata dal modello. Il sistema di ipotesi è ancora $H_0 : R^2 = 0$ contro $H_1 : R^2 > 0$, che in realtà coincide con il verificare l'ipotesi nulla che tutte le pendenze del modello siano uguali a zero. Nella regressione semplice vi era solo una pendenza, perciò il test coincideva con il test t di significatività di β_1; nella regressione multipla, invece, si tratta di una sorta di test globale sulla bontà (cioè l'utilità) del modello nello spiegare la variabilità della **y**.

Per calcolarlo sul campione, di solito ci si avvale della formula:

$$\hat{R}^2 = 1 - \frac{\text{Dev}(e)}{\text{Dev}(\mathbf{y})}$$

La devianza degli errori del campione è ottenibile come:

$$\text{Dev}(e) = (n-p)\, s_\varepsilon^2$$

Nel nostro caso s_ε è il `Residual standard error`, uguale a 83.07. Quindi:

$$\text{Dev}(e) = (5-3)\, 83.07^2 = 13800.56$$

Sapendo, dall'esercizio sulla regressione semplice, che $\text{Dev}(\mathbf{y}) = 160600.4$, possiamo stimare l'$R^2$:

$$\hat{R}^2 = 1 - \frac{13800.56}{160600.4} = 0.9141$$

La statistica test (e la sua distribuzione sotto H_0) è:

$$T_n = \frac{\left(\frac{\hat{R}^2}{p-1}\right)}{\left(\frac{1-\hat{R}^2}{n-p}\right)} \sim \mathbf{F}_{p-1,\,n-p}$$

Verificare la bontà del modello usando $\alpha = 10\%$.

L'area di non rifiuto di H_0 è dunque:

$$\mathcal{A}_d = [0,\, F_{0.1,\,3-1,\,5-3}] = [0,\, 9]$$

Il valore osservato della statistica test è riportato nell'output del software come `F-statistic` ed è 10.64, infatti:

$$F_{\text{oss}} = \frac{\left(\frac{0.9141}{3-1}\right)}{\left(\frac{1-0.9141}{5-3}\right)} = 10.64 \notin A_d \longrightarrow \text{rifiuto } H_0 \text{ (infatti il } p\text{-value nell'output è } 0.08593 < 0.1)$$

Possiamo quindi concludere che, per un livello di significatività del 10%, vi è un contributo congiunto significativo delle variabili indipendenti nello spiegare la variabile di risposta.

3 ANOVA ad una via con soggetti indipendenti

L'ANOVA (**AN**alysis **O**f **VA**riance) non è altro che un'estensione del test t di confronto tra medie, dove il numero di gruppi G è maggiore o uguale a due (nel caso di 2 gruppi, si avranno risultati equivalenti al test t con stimatore congiunto per la varianza). Considereremo lo scenario in cui i G campioni – ognuno con una propria numerosità n_j ($j = 1, \ldots, G$) – provengano da popolazioni indipendenti X_j, ciascuna Normale con una certa media μ_j ed una varianza σ^2 incognita ma che assumiamo comune a tutte le popolazioni. Considerando nel nostro caso un solo fattore di stratificazione in gruppi, il modello prende il nome di "ANOVA ad una via", traduzione sin troppo letterale della denominazione inglese "one-way ANOVA".

Date G popolazioni, il sistema di ipotesi è il seguente:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_G$$

$$H_1 : \text{ almeno una } \mu_j \text{ è diversa dalle altre}$$

La statistica test è il rapporto tra due stimatori di σ^2:

- al numeratore: $s^2_{\text{est}} = \frac{\text{Dev}_{\text{est}}}{G-1} = \frac{\sum_{j=1}^{G} (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^2\, n_j}{G-1}$, valido sotto H_0;
- al denominatore: $s^2_{\text{int}} = \frac{\text{Dev}_{\text{int}}}{n-G} = \frac{\sum_{j=1}^{G} s_j^2\,(n_j - 1)}{n-G}$, valido sotto entrambe le ipotesi.

Essendo la statistica test un rapporto tra due distribuzioni χ^2, si distribuisce come una $\mathbf{F}$ di Fisher con $G-1$ e $n-G$ gradi di libertà. Se l'ipotesi nulla è vera, $\sigma^2_{\text{est}} = \sigma^2_{\text{int}}$, cioè il rapporto deve essere uguale a 1. In caso H_0 non sia vera, allora s^2_{est} tenderà a sovrastimare la varianza. Per questo motivo l'ipotesi alternativa è quella di un test unidirezionale destro in cui il rapporto tra le σ^2_{est} e σ^2_{int} sia maggiore di 1.

Facciamo un esercizio: vi sono tre gruppi di studenti (A, B, C) che svolgono un test. Si è interessati ad eventuali differenze in punteggio medio tra i tre gruppi sapendo che:

- $n_A = 30,\ n_B = 14,\ n_C = 20$;
- $\text{Dev}_A = 5572.345,\ \text{Dev}_B = 1489.162,\ \text{Dev}_C = 3811.824$;
- $\bar{\mathbf{x}}_A = 61.237,\ \bar{\mathbf{x}}_B = 63.189,\ \bar{\mathbf{x}}_C = 72.765$.

Completare la seguente tabella ed eseguire il test ANOVA usando $\alpha = 0.01$.

Fonte di variabilità	Devianze	gradi di libertà	Varianze	F_{oss}
Esterna Interna				
Totale				

Svolgimento:

Prima di tutto calcoliamo la media generale:

$$\bar{\mathbf{x}} = \frac{30 \times 61.236 + 14 \times 63.189 + 20 \times 72.765}{64} = 65.267$$

- II colonna tabella:
  - La devianza esterna è la somma pesata degli scostamenti al quadrato delle medie di gruppo dalla media generale:

$$(61.237 - 65.267)^2 \times 30 + (63.189 - 65.267)^2 \times 14 + (72.765 - 65.267)^2 \times 20 = 1672.18$$

  - La devianza interna è la somma delle devianze di gruppo:

$$\text{Dev}_{\text{int}} = 5572.345 + 1489.162 + 3811.824 = 10873.33$$

  - La devianza totale è la somma delle due devianze (verificare che sia uguale alla devianza di $\mathbf{x}$):

$$\text{Dev}_{\text{tot}} = \text{Dev}_{\text{est}} + \text{Dev}_{\text{int}} = 12545.51$$

- III colonna tabella:
  - gradi di libertà del numeratore: $3 - 1 = 2$;
  - gradi di libertà del denominatore: $64 - 3 = 61$;
  - gradi di libertà totali: $64 - 1 = 63$
- IV colonna tabella:
  - La varianza esterna è uguale a:

$$s^2_{\text{est}} = \frac{1672.18}{2} = 836.09$$

  - La varianza interna è uguale a:

$$s^2_{\text{int}} = \frac{10873.33}{61} = 178.25$$

  - La varianza totale è uguale a:

$$s^2_{\text{tot}} = \frac{12545.51}{63} = 199.14$$

- V colonna tabella:

  Il valore osservato della statistica $\mathbf{F}$ è uguale a:

$$F_{\text{oss}} = \frac{836.09}{178.25} = 4.69$$

La regione di accettazione unidirezionale destra è $\mathcal{A}_d = [0,\, F_{0.01,\,2,\,61}] = [0,\, 4.97]$, che contiene F_{oss}. Dunque non rifiutiamo H_0.

L'output del software è infatti:

```
##             Df  Sum Sq Mean Sq F value  Pr(>F)
## gruppi       2  1672.2  836.09  4.6905 0.01274 *
## Residuals   61 10873.3  178.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

dove il p-value è maggiore di 0.01, quindi non rifiutiamo l'ipotesi nulla.

4 Test non parametrici

I test non parametrici non necessitano di ipotesi a priori sulle caratteristiche della popolazione (ovvero, di un parametro), o comunque le ipotesi sono meno restrittive di quelle usate nei test parametrici.

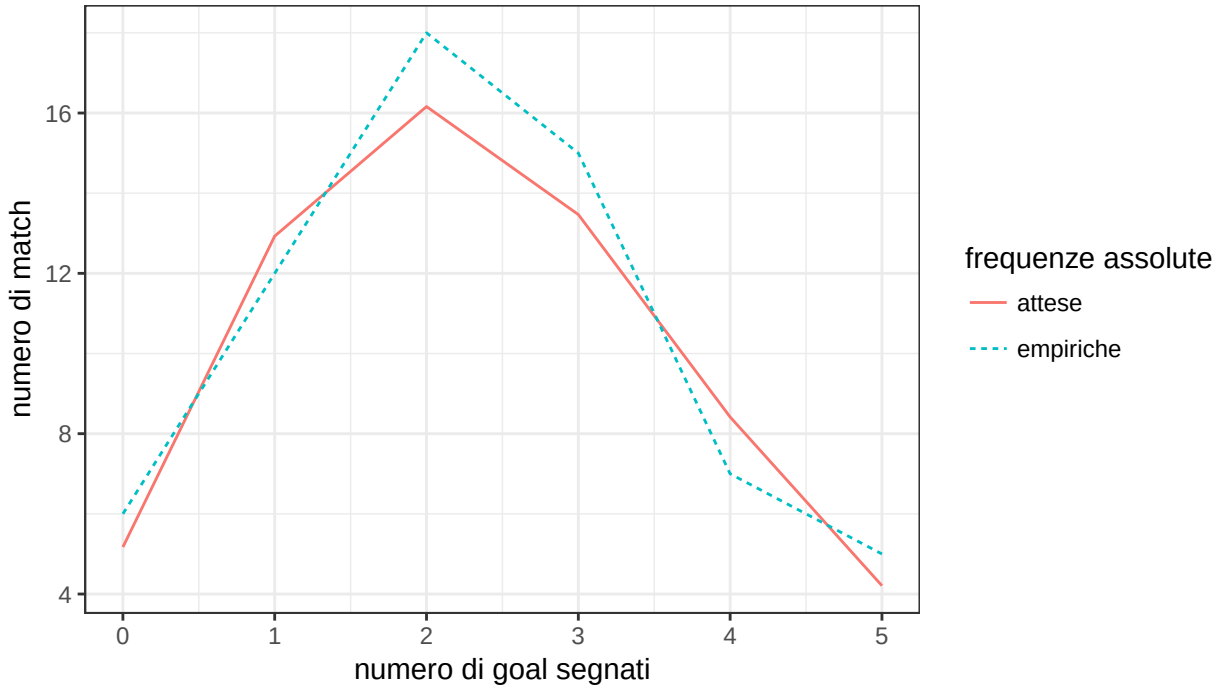
4.1 Test di bontà di adattamento

Il test di bontà di adattamento verifica l'ipotesi nulla che una tabella di frequenze per la variabile discreta X segua una determinata distribuzione. Ragiona sul fatto che, nel caso in cui la distribuzione assunta sotto H_0 sia vera, le frequenze attese sotto tale distribuzione non dovrebbero differire in maniera sostanziale da quelle riscontrate empiricamente. Il confronto tra frequenze attese e frequenze empiriche è oggetto di un test la cui statistica segue una distribuzione χ^2 con $G-1$ gradi di libertà, dove G rappresenta il numero di modalità della variabile X.

Prendiamo per esempio una competizione calcistica in cui si sono giocati $N = 63$ match totali, tutti di 90 minuti. Vi sono state $n_0 = 6$ partite con zero goal, $n_1 = 12$ partite con un goal, $n_2 = 18$ partite con due gol, $n_3 = 15$ partite con tre gol, $n_4 = 7$ con quattro goal e $n_5 = 5$ con cinque goal. Quindi vi sono $G = 6$ modalità per la variabile X "numero di goal segnati in un match".
Verificare l'ipotesi nulla che X segua un modello di Poisson di parametro $\lambda = 2.5$, utilizzando $\alpha = 10\%$.

Per far ciò, bisogna:

1. Calcolare le frequenze assolute attese sotto l'ipotesi nulla $H_0 : X \sim \mathbf{Poi}(2.5)$:
   - $\hat{n}_0 = N \times \mathbf{P}(k=0) = N \times \frac{\lambda^k e^{-\lambda}}{k!} = 63 \times e^{-2.5} = 5.171$
   - $\hat{n}_1 = N \times \mathbf{P}(k=1) = 63 \times 2.5\, e^{-2.5} = 12.928$
   - $\hat{n}_2 = N \times \mathbf{P}(k=2) = 63 \times \frac{2.5^2\, e^{-2.5}}{2} = 16.160$
   - Allo stesso modo, risulta $\hat{n}_3 = 13.467,\ \hat{n}_4 = 8.417,\ \hat{n}_5 = 4.208$ (verificare ciò come esercizio).
2. Determinare la regione di non rifiuto di H_0:
$$\mathcal{A}_d = [0,\, \chi^2_{0.1,\,6-1}] = [0,\, 9.236]$$
3. Calcolare il valore osservato della statistica χ^2 sommando le distanze quadratiche relative tra ciascuna frequenza empirica e la corrispondente frequenza attesa:
$$\chi^2_{\text{oss}} = \sum_{k=0}^{G-1} \frac{(n_k - \hat{n}_k)^2}{\hat{n}_k} = \frac{(6-5.171)^2}{5.171} + \frac{(12-12.928)^2}{12.928} + \ldots + \frac{(5-4.208)^2}{4.208} = 0.971$$
4. $\chi^2_{\text{oss}} \in \mathcal{A}_d \longrightarrow$ non rifiuto H_0.



4.2 Test di indipendenza assoluta

Il test di indipendenza assoluta χ^2 verifica l'ipotesi nulla che non vi sia connessione tra due variabili qualitative. Non ci stiamo chiedendo se una in particolare influenza l'altra, ma semplicemente se vi è un'interdipendenza tra le due. Abbiamo già visto l'indice χ^2 in Statistica descrittiva, ora aggiungeremo un ulteriore step, quello inferenziale, per avere informazioni sulla connessione tra le due variabili nella popolazione partendo dall'esperimento campionario.

Una piccola azienda farmaceutica è interessata a capire se vi sia connessione tra l'acquisto di determinate categorie di medicinali e le diverse fasce di età dei pazienti. Di seguito, la tabella che riassume i prodotti acquistati in media ogni giorno per categoria di medicinale e per fascia di età.

	Bambini	Giovani	Adulti	Anziani	TOT di riga
Antistaminici	110	80	110	120	420
Antidolorifici	40	50	200	300	590
Antibiotici	250	120	150	350	870
TOT di colonna	400	250	460	770	1880

Verificare l'ipotesi di non connessione tra le due variabili utilizzando $\alpha = 1\%$.

- Calcoliamo la tabella delle frequenze assolute attese sotto l'ipotesi di indipendenza. Ricordiamo che, sotto tale ipotesi, $\hat{n}_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n_{\bullet\bullet}}$:

	Bambini	Giovani	Adulti	Anziani	TOT di riga
Antistaminici	89.362	55.851	102.766	172.021	420
Antidolorifici	125.532	78.457	144.362	241.649	590
Antibiotici	185.106	115.692	212.872	356.330	870
TOT di colonna	400	250	460	770	1880

- Il sistema di ipotesi è:
  - $H_0 : \chi^2 = 0$
  - $H_1 : \chi^2 > 0$
- La regione di non rifiuto di H_0 è:

$$\mathcal{A}_d = \left[0,\, \chi^2_{0.01,\underbrace{(3-1)}_{R-1} \times \underbrace{(4-1)}_{C-1}} \right] = [0,\, 16.812]$$

- Il valore osservato della statistica è:

$$\chi^2_{\text{oss}} = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = 177.175$$

- $\chi^2_{\text{oss}} \notin \mathcal{A}_d \longrightarrow$ rifiuto H_0.