CALCOLATORE DI REGRESSIONE LINEARE

Metodi Matematici e Statistici, A.A. 2019/2020

Simone Torrisi X81000732

Link al progetto: http://torrisisimone.altervista.org/regrlin/index.html

INDICE

INTRODUZIONE

Il progetto è incentrato sullo sviluppo di un'applicazione web relativa ad un calcolatore di regressione lineare. Acquisito un dataset inserito dall'utente, l'obiettivo è trovare la retta che meglio approssima i dati, cercando una correlazione tra essi.

Il software è in grado di:

- Prendere i valori **x** e **y** inseriti dall'utente;
- Calcolare i coefficienti **m** e **q** della retta di regressione;
- Calcolare la **covarianza** e il **coefficiente di correlazione lineare** (o di Pearson);
- Calcolare gli intervalli di confidenza per i coefficienti della retta, andando ad individuare una **banda di confidenza**;
- Effettuare delle **stime** di nuovi valori sulla base della serie di dati precedente;
- Fornire un risultato grafico.

REGRESSIONE LINEARE

La regressione lineare rappresenta un metodo di stima del valore atteso partendo da due variabili x e y legati da una retta.

Tale retta è del tipo: $\boldsymbol{y = mx + q}$

Dove **q** è l'intercetta, ovvero il punto dove la retta incontra l'asse delle ordinate, mentre **m** è il coefficiente angolare della retta che indica di quanto aumenta y all'aumentare di un'unità di x.

La retta di regressione si ottiene applicando il metodo dei minimi quadrati.

METODO DEI MINIMI QUADRATI

Tecnica che permette di trovare una funzione che minimizza lo scarto quadratico tra un generico punto y e la retta di regressione.

La funzione è: $g(f) = \sum_{i=1}^{n} |\mathrm{f}(\mathrm{x_i}) - y_i|^2$

Per trovare i coefficienti della retta di regressione si applicano le variabili m e q alla funzione e diventa: $\boldsymbol{g(m,q) = \sum_{i=1}^{n} [mx_i + q - y_i]^2}$

Risolvendo il sistema e usando le formule della varianza e covarianza, si trova che:

$$\boldsymbol{m = \frac{c_{xy}}{s_x^2}} \qquad \boldsymbol{q = \bar{y} - \frac{c_{xy}}{s_x^2}\,\bar{x}}$$

VARIANZA E DEVIAZIONE STANDARD

La **varianza** è un indice che misura di quanto i dati si discostano dal valore medio. Tanto più è grande questo valore, tanto più i dati sono "distanti" dalla media. Se è nullo, tutti i dati sono uguali.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

La **deviazione standard** (o scarto quadratico medio) è un indice risultante dalla radice quadrata della varianza.

$$s = \sqrt{s^2}$$

COVARIANZA

La **covarianza** è un indice che esprime una correlazione tra la variabile x e la variabile y.

$$c_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})$$

Due serie di dati sono statisticamente incorrelate se $c_{xy} = 0$. Tuttavia non ci assicura che non ci sia una dipendenza tra le due serie di dati.

Se $c_{xy} > 0$ i due set di dati si dicono correlati positivamente, se $c_{xy} < 0$ i due set di dati si dicono correlati negativamente.

COEFFICIENTE DI PEARSON

Il **coefficiente di correlazione lineare** (o di Pearson) è un indice che esprime un'eventuale relazione di linearità tra due serie di dati.

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$

È compreso nell'intervallo [-1,1]. Più il valore si avvicina ad uno degli estremi, più i dati sono allineati con la retta di regressione.

Se $r_{xy} > 0$ la retta è ascendente, se $r_{xy} < 0$ la retta è discendente. Se i numeri sono piccoli può capitare che la covarianza sia molto vicina a zero, ma che il coefficiente di Pearson sia molto vicino ad uno degli estremi.

BANDA DI CONFIDENZA

Fissato α ∈ [0,1], si chiama **intervallo di confidenza** con livello di fiducia α, quell'intervallo: $[T - e_1\ ,\ T + e_2]$ tale che $P(T' \in\ [T - e_1\ ,\ T + e_2]) = 1 - \alpha$.

L'obiettivo è trovare gli intervalli di confidenza dei coefficienti m e q in modo da calcolare la **banda** (o striscia) **di confidenza**.

Siano $[m - m_1\ ,\ m + m_1]$ e $[q - q_1\ ,\ q + q_1]$ gli intervalli di confidenza per i coefficienti m e q della retta.

$$\boldsymbol{m_1 = t_{1-\frac{\alpha}{2}} \sqrt{s_{RES}^2\ \frac{1}{s_x^2}}} \qquad \boldsymbol{q_1 = t_{1-\frac{\alpha}{2}} \sqrt{s_{RES}^2\ [\frac{1}{n} + \frac{\bar{x}^2}{s_x^2}]}}$$

Dove $s_{RES}^2 = \frac{1}{n-2}\ \sum_{i=1}^{n}(y_i - y_i')^2$, con y_i' si intendono i valori teorici che dovrebbero essere assunti dal valore y se si trovassero sulla retta (in assenza di errori). Questo indice è chiamato **Errore Standard**, quanto più è piccolo tanto più attendibile è il valore statistico calcolato. Il termine $t_{1-\frac{\alpha}{2}}$ indica il quantile di ordine $1 - \frac{\alpha}{2}$ della distribuzione t-student con n-2 gradi di libertà.

La banda di confidenza sarà compresa tra le rette:

$$\boldsymbol{y_1 = (m - m_1)\ x + (q - q_1)} \qquad \boldsymbol{y_2 = (m + m_1)\ x + (q + q_1)}$$

STIMA DEI VALORI

È possibile stimare il valore di y, dato in input un determinato valore x o viceversa.

In presenza di una banda di confidenza, al valore stimato corrisponderà un intervallo entro il quale si stima che cada, con una certa probabilità basata su un valore di α scelto, il valore vero.

Siano [mMin, mMax] e [qMin, qMax] gli intervalli di confidenza di m e q trovati in precedenza e sia x_1 il valore di cui si vuole trovare la stima del corrispettivo y, l'intervallo stimato è dato da: $\boldsymbol{[mMin \times x_1 + qMin\ ,\ mMax \times x_1 + qMax]}$

Per stimare il valore di x, dato in input un determinato valore di y, l'intervallo risultante è dato dalle formule inverse: $\boldsymbol{[\frac{y_1 - qMin}{mMin}\ ,\ \frac{y_1 - qMax}{mMax}]}$

IMPLEMENTAZIONE PRATICA

Di seguito è mostrato il codice delle funzioni per il calcolo di tutti i coefficienti.

L'implementazione fa riferimento ai concetti e alle formule teoriche precedenti.

Il linguaggio utilizzato è Javascript.

VARIANZA

```
calcolaVarianzaX() {
        const { x, mediaX } = this.state;
        let n = x.length;
        let sommatore = 0;
        for (let i = 0; i < n; i++) {
            sommatore += Math.pow((x[i] - mediaX), 2);
        }
        varianzaX = sommatore / n;

calcolaVarianzaY() {
        const { y, mediaY } = this.state;
        let n = y.length;
        let sommatore = 0;
        for (let i = 0; i < n; i++) {
            sommatore += Math.pow((y[i] - mediaY), 2);
        }
        varianzaY = sommatore / n;
    }
```

DEVIAZIONE STANDARD

```
calcolaDevstdX() {
        const { varianzaX } = this.state;
        devstdX = Math.sqrt(varianzaX);
   }
calcolaDevstdY() {
        const { varianzaY } = this.state;
        devstdY = Math.sqrt(varianzaY);
    }
```

COVARIANZA

```
calcolaCovarianza() {
        const { x, y, mediaX, mediaY } = this.state;
        let n = x.length;
        let sommatore = 0;
        for (let i = 0; i < n; i++) {
            sommatore += (x[i] - mediaX) * (y[i] - mediaY);
        }
        covarianza = sommatore / n;
    }
```

COEFFICIENTE DI PEARSON

```
calcolaPearson() {
        const { covarianza, devstdX, devstdY } = this.state;
        pearson = covarianza / (devstdX * devstdY);
    }
```

COEFFICIENTI DELLA RETTA

```
calcolaCoeffM() {
        const { covarianza, varianzaX } = this.state;
        m = covarianza / varianzaX;
    }
calcolaCoeffQ() {
        const { mediaX, mediaY, covarianza, varianzaX } = this.state;
        q = mediaY - (mediaX * covarianza / varianzaX);
    }
```

QUANTILE DELLA T-STUDENT

```
calcolaQuantile() {
        const { x, alfa } = this.state;
        var { jStat } = require('jstat');
        var ordine = 1 - (alfa / 2);
        quantile = jStat.studentt.inv(ordine, x.length - 2);
    }
```

ERRORE STANDARD

```
calcolaS2RES(stimeY) {
        const { y } = this.state;
        let n = y.length;
        let sommatore = 0;
        for (let i = 0; i < n; i++) {
            sommatore += Math.pow((y[i] - stimeY[i]), 2);
        }
        s2res = sommatore / (n - 2);
    }
```

BANDA DI CONFIDENZA

```
calcolaIntervallo() {
        const{ x, quantile, mediaX,
              coeffM, coeffQ, s2res } = this.state;
        let mMin, mMax, qMin, qMax;
        let n = x.length;

        m1 = quantile * Math.sqrt(s2RES / varianzaX);
        mMin = coeffM - m1;
        mMax = coeffM + m1;

       q1 = quantile*Math.sqrt(s2RES*(1/n+Math.pow(mediaX,2)/varianzaX));
        qMin = coeffQ - q1;
        qMax = coeffQ + q1;
    }
```

STIMA DEI VALORI

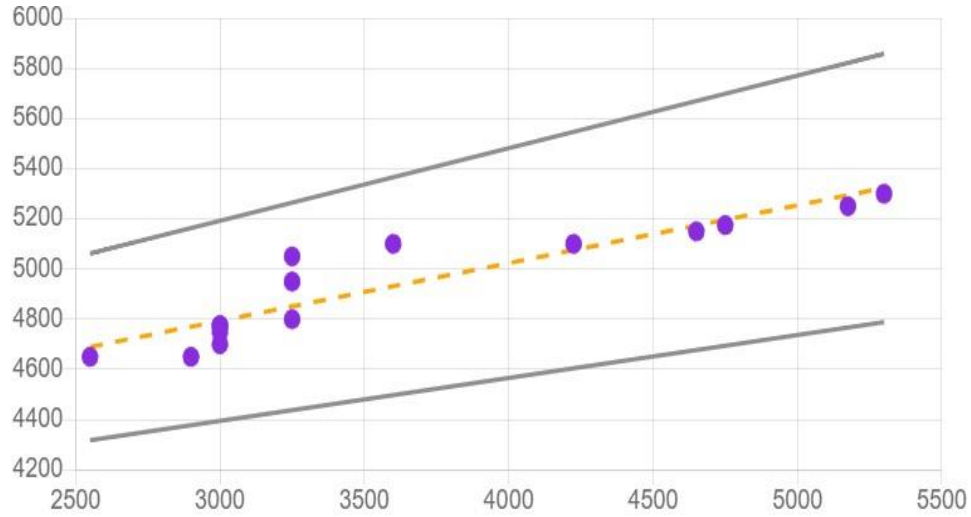
```
    calcolaStimaX() {
        const { stimaY, mMin, mMax, qMin, qMax } = this.state;

        stimaXMin = (stimaY - qMin) / mMin;
        stimaXMax = (stimaY - qMax) / mMax;
    }

    calcolaStimaY() {
        const { stimaX, mMin, mMax, qMin, qMax } = this.state;

        stimaYMin = mMin * stimaX + qMin;
        stimaYMax = mMax * stimaX + qMax;
    }
```

RISULTATO GRAFICO

Di seguito è proposto il risultato grafico di un set di dati di prova (tabella a lato). La prima colonna costituisce la variabile x, la seconda costituisce la variabile y.
Nel grafico i punti viola rappresentano le coppie della tabella, la retta gialla tratteggiata rappresenta la retta di regressione lineare che meglio approssima i dati, le rette di colore grigio indicano la banda di confidenza con $\alpha = 0{,}05$ (95%).

Tabella 1.5: Carichi di rottura

I^a lesione	rottura
2550	4650
2900	4650
3000	4700
3000	4750
3000	4775
3000	4775
3250	4800
3250	4950
3250	5050
3600	5100
4225	5100
4650	5150
4750	5175
5175	5250
5300	5300



VALORI NUMERICI

$m = 0{,}230$ $\quad m_1 = 0{,}059$ $\quad m: [0{,}171\,;\,0{,}289]$

$q = 4101{,}746$ $\quad q_1 = 222{,}068$ $\quad q: [3879{,}678\,;\,4323{,}814]$

$c_{xy} = 177133{,}333$

$r_{xy} = 0{,}920$

REFERENZE

Dimostrazione formule dei coefficienti m e q:

- Orazio Muscato, (2019). Metodi Matematici e Statistici. Regressione lineare (p. 33)

Documentazione formule intervallo di confidenza dei coefficienti m e q:

- Francesco Lagona, Intervalli di confidenza. Varianza incognita (p. 4)

Documentazione libreria Javascript JStat per il calcolo del quantile della t-student:

- https://github.com/jstat/jstat

Documentazione React, framework Javascript utilizzato per lo sviluppo del progetto:

- https://it.reactjs.org/