**Relazione tecnica**

*Francesca Behrens, Paolo Botta, Romeo Carrara*

**Presentazione del problema**

L’azienda W&H è produttore a livello mondiale di apparecchiature di precisione in campo odontoiatrico. In particolare la filiale italiana W&H Sterilization con sede a Brusaporto (BG) si occupa della produzione di macchine sterilizzatrici. Ogni giorno i dentisti di tutto il mondo usufruiscono di queste macchine per sterilizzare le attrezzature che utilizzano sui loro pazienti.

Il ciclo o processo di sterilizzazione si articola in varie fasi denominate PV1, PP1, PV2, PP2, PV3, PPH1, PPH2 (per ulteriori informazioni si veda l’approfondimento tecnico in allegato) di durata variabile. Per sterilizzare il carico viene fatto uso di vapore e di conseguenza le apparecchiature rimangono bagnate. Occorre quindi, alla fine del ciclo, asciugare il carico.

Come possiamo facilmente intuire il tempo di asciugatura dipende dagli oggetti presenti all’interno della sterilizzatrice. Molteplici osservazioni effettuate in passato hanno evidenziato che il tempo di asciugatura dipende solo ed esclusivamente dal peso del carico. La macchina, però, non possiede una bilancia.

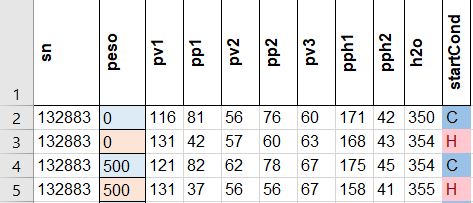
Gli statistici devono quindi stimare il peso delle apparecchiature basandosi sugli altri dati che la macchina è in grado di misurare, nello specifico la durata delle fasi PV1, PP1, PV2, PP2, PV3, PPH1, PPH2 e la quantità di acqua utilizzata.

**Presentazione dei dati**

Il laboratorio dell’azienda ha svolto numerose misurazioni e i dati raccolti di interesse per la statistica sono stati sintetizzati nel dataset “*VA131.txt”*.

* La prima colonna del file contiene il numero seriale della macchina oggetto di misurazione.
* La seconda colonna contiene il peso del carico inserito (in grammi).
* La colonna “*h2o*” indica il volume di acqua utilizzato per il ciclo (in centilitri).
* La colonna “*startCond*” indica in quali condizioni iniziali è stato svolto l’esperimento: freddo (C, *cold*) oppure caldo (H, *hot*).
* Le colonne rimanenti contengono la durata (in secondi) della fase alla quale fanno riferimento.

I test sono stati svolti su 8 differenti macchine dello stesso modello. Per ogni macchina sono stati svolti dei cicli con carichi da 13 pesi diversi (0-6 Kg) nelle due *varianti* Hot e *Cold*. In totale si hanno quindi a disposizione 208 cicli da analizzare.



Dataset di lavoro

**Analisi**

Per prima cosa occorre capire se le condizioni iniziali (*Hot* o *Cold*) influenzano significativamente i valori dei dati. Possiamo farlo sia qualitativamente, attraverso un’esplorazione grafica (riportata in allegato), sia quantitativamente, attraverso un test d’ipotesi.

Sia PV1, PP1, PV2, PP2, PV3, PPH1, PPH2 e H2O la lista dei possibili predittori del peso del carico.  
Sia ora X un generico predittore tra quelli appena elencati.  
Si definisca la variabile aleatoria che assume come valori la differenza tra il valore di X misurato in condizioni iniziali “Cold” e in condizioni “Hot” a parità di carico e dalla stessa macchina.

Se la media reale di è diversa da 0 è ragionevole supporre che le condizioni iniziali influenzino i valori delle grandezze considerate. Impostiamo quindi il test statistico:

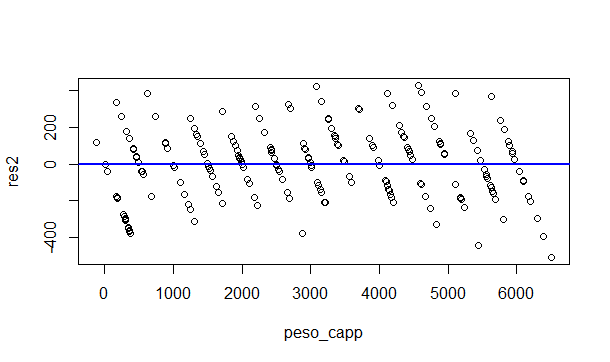
|  |  |
| --- | --- |
| Predittore | p-value |
| PV1 | < 2.2e-16 |
| PP1 | < 2.2e-16 |
| PV2 | < 2.2e-16 |
| PP2 | < 2.2e-16 |
| PV3 | < 2.2e-16 |
| PPH1 | < 2.2e-16 |
| PPH2 | < 2.2e-16 |
| H2O | 2.553e-10 |

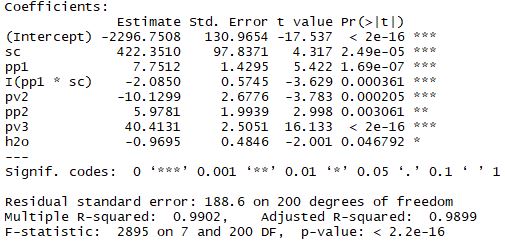
Per il T.C.L, essendo un campione numeroso,

Avendo a disposizione molte misurazioni (104) si approssima la varianza con la varianza campionaria.

Dai valori dei p-value estremamente bassi si può facilmente concludere che le condizioni iniziali sono di fondamentale importanza e, per la corretta stima del peso, bisognerà anche considerare un predittore categorico (denominato *sc*).

Dopo varie simulazioni con R (per approfondimenti si veda il file R in allegato) e alcune osservazioni grafiche si conclude che il modello lineare migliore (e contemporaneamente più snello) per la stima del peso risulta essere:

I valori dei regressori sono riportati nello screenshot del codice R sottostante. Visti i p-value molto bassi possiamo affermare che tutti i regressori individuati sono diversi da 0 e quindi molto significativi. Gli indici R e R-adjusted sono molto alti. I residui sono inoltre omoschedastici (come mostra il grafico) e distribuiti normalmente (shapiro test: 0.4472).



Codice R sulla significatività dei regressori

Grafico dei residui

**Conclusioni**

Attraverso un opportuno test statistico si è verificato che le condizioni iniziali influenzano significativamente i valori delle varie grandezze. Per stimare il peso delle apparecchiature è quindi necessario prendere in considerazione un predittore categorico. È stato infine costruito un modello di regressione lineare multipla (molto valido) per la stima del peso medio del carico.

A fini pratici sarebbe utile costruire un intervallo di predizione per un’osservazione futura ma, trattandosi di regressione multipla, non è possibile trovare una formula analitica e al massimo si possono fornire stime numeriche tramite un software. Per un esempio si veda il file R in allegato.