R per l’analisi statistica multivariata

Unità F: analisi descrittiva dei dati FORBES

Tommaso Rigon

Università Milano-Bicocca

Unità F

Argomenti affrontati

- Svolgimento di un tema d'esame di Statistica I
- Modello di regressione lineare semplice
- Modelli linearizzabili

- Esercizi **R** associati: `https://tommasorigon.github.io/introR/exe/es_2.html`

Descrizione del problema

- Per $n = 17$ luoghi nelle Alpi viene misurata la **pressione atmosferica** (inHg, ovvero "inches of mercury") e la **temperatura di ebollizione** dell'acqua (in gradi Fahrenheit).
- I dati provengono da un esperimento condotto dal **fisico scozzese Forbes** nel 1857.
- Forbes era interessato a stimare l'altitudine tramite la pressione. Tuttavia, il barometro all'epoca era uno strumento pesante e costoso.
- In montagna infatti l'acqua bolle ad una temperatura diversa, per cui è possibile cercare di stimare la pressione a partire dalla temperatura di ebollizione.
- **Nota**. I dati seguenti sono gli stessi dell'esame di Statistica I del 11 Novembre 2020, che trovate sul sito web.

I dati grezzi (editor di testo)

```
"bp","pres"
194.5,20.79
194.3,20.79
197.9,22.4
198.4,22.67
199.4,23.15
199.9,23.35
200.9,23.89
201.1,23.99
201.4,24.02
201.3,24.01
203.6,25.14
204.6,26.57
209.5,28.49
208.6,27.76
```

Importazione dei dati forbes

- Come fatto in precedenza, anzitutto è necessario scaricare il file `forbes.csv` e salvarlo nel proprio computer.
- **Link al file**: `https://tommasorigon.github.io/introR/data/forbes.csv`.
- In alternativa, possiamo scaricare il file direttamente da internet nel modo seguente:

```
forbes <- read.table("https://tommasorigon.github.io/introR/data/forbes.csv",
                     header = TRUE, sep = ",")
str(forbes)

# 'data.frame':       17 obs. of  2 variables:
#  $ bp  : num  194 194 198 198 199 ...
#  $ pres: num  20.8 20.8 22.4 22.7 23.1 ...
```

Operazioni preliminari

- Per motivi **interpretativi**, convertiamo la temperatura da gradi Farenheit a gradi Celsius, ricordando che

$$(\text{“Fahrenheit”}) = 32 + \frac{9}{5}(\text{“Celsius”}).$$

```r
colnames(forbes) <- c("TempF", "Pressione") # Cambio i nomi alle variabili

forbes$TempC <- round((forbes$TempF - 32) * 5 / 9, 2) # Da Fahrenheit a Celsius
summary(forbes)
#      TempF         Pressione          TempC
#  Min.   :194.3   Min.   :20.79   Min.   : 90.17
#  1st Qu.:199.4   1st Qu.:23.15   1st Qu.: 93.00
#  Median :201.3   Median :24.01   Median : 94.06
#  Mean   :203.0   Mean   :25.06   Mean   : 94.97
#  3rd Qu.:208.6   3rd Qu.:27.76   3rd Qu.: 98.11
#  Max.   :212.2   Max.   :30.06   Max.   :100.11
```

- Come mai approssimiamo i valori utilizzando `round`? Per motivi **estetici**: in questo modo si ottengono risultati identici alla prova d'esame di Statistica I.

Domanda I

Domanda

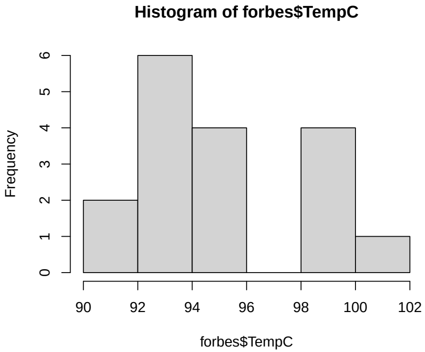
Si disegni un istogramma della variabile temperatura, scegliendo un numero appropriato di classi equispaziate e giustificandone la scelta.

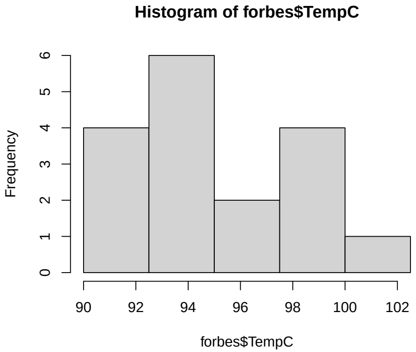
- Possiamo decidere di specificare in autonomia gli intervalli delle classi oppure di lasciare ad **R** questa scelta.

```r
par(mfrow = c(1, 2)) # Divido la finestra grafica in 2 parti

# Opzione 1, per un totale di 6 classi equispaziate
hist(forbes$TempC) # Equivalente a: hist(forbes$TempC, breaks = "sturges")

# Opzione 2, definisco manualmente 5 classi equispaziate
breaks <- c(90, 92.5, 95, 97.5, 100, 102.5)
hist(forbes$TempC, breaks = breaks)
```

- **Esercizio**. Si aggiungano a questi grafici gli “abbellimenti” grafici ritenuti necessari (nomi delle variabili, titolo, etc).

Domanda I





- La soluzione di sinistra fa uso di 6 classi. Viceversa, quella di sinistra fa uso di 5 classi, come nella **soluzione** dell'esame.

Domanda II

Domanda

Si ottengano la media aritmetica di entrambe le variabili `tempC` e `pressione`. Quanto vale la temperatura di ebollizione media espressa in gradi Fahrenheit? Si risponda senza calcolare tutti i valori della variabile `tempF`.

- Abbiamo già calcolato la medie tramite il comando `summary`, per completezza:

```
# Prima parte della domanda
mean(forbes$TempC)
# [1] 94.97353
mean(forbes$Pressione)
# [1] 25.05882

# Seconda parte della domanda
32 + 9 / 5 * mean(forbes$TempC) # Utilizzo proprietà della media
# [1] 202.9524
mean(forbes$TempF) # Non richiesto, calcola la media a partire dai dai dati trasformati
# [1] 202.9529
```

Esercizio. Come mai le medie calcolate nella seconda parte differiscono leggermente? A cosa può essere dovuto?

Domanda III

Domanda

Si ottenga la varianza delle variabili `tempC` e `pressione`.

- Dato che tornerà utile in seguito, definiamo la funzione `my_var` che calcola la **varianza**.

```
# Si, la funzione è definita in un'unica riga e non c'è nulla di male in questo
my_var <- function(x) mean(x^2) - mean(x)^2

# Calcolo delle due varianze
my_var(forbes$TempC)
# [1] 9.630811
my_var(forbes$Pressione)
# [1] 8.584575
```

- **Esercizio**. Si ottengano i momenti secondi delle variabili `tempC` e `pressione`.

Domanda IV

Domanda

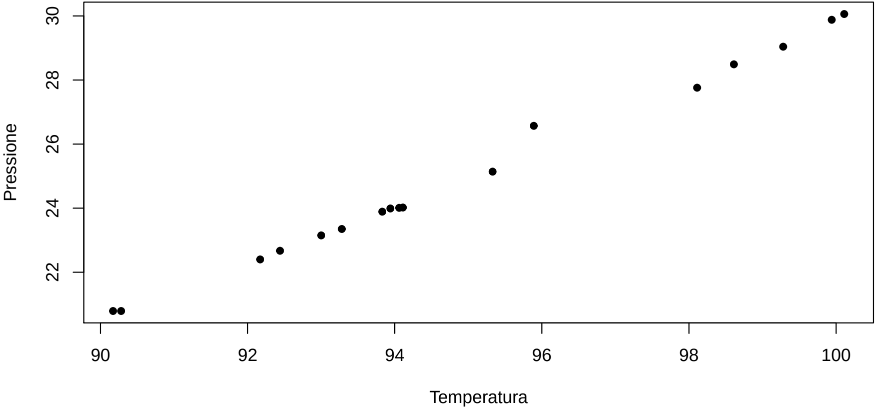
Si disegni un opportuno grafico che aiuti a comprendere la relazione tra le due variabili. Si calcoli quindi la correlazione.

- Per la prima parte della domanda, abbiamo bisogno del nuovo comando `plot`, che può essere usato (tra le altre cose!) per costruire un **diagramma a dispersione**.
- Forniamo due versioni dello stesso grafico; la seconda contiene dei miglioramenti estetici.

```r
par(mfrow = c(1, 1)) # Vogliamo mostrare un grafico alla volta
plot(forbes$TempC, forbes$Pressione)
plot(forbes$TempC, forbes$Pressione, pch = 16, xlab = "Temperatura", ylab = "Pressione")
```

- È quindi evidente che i dati siano circa (anche se non perfettamente) allineati

Domanda IV



Domanda IV

- Per la seconda parte di domanda (correlazione), dobbiamo anzitutto ottenere la **covarianza** tra due variabili.

- La **covarianza** tra due insiemi di dati $x_1, \dots, x_n$ e $y_1, \dots, y_n$ è definita come

$$\text{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

- Definiamo quindi la funzione `my_cov`, che calcola appunto la covarianza:

```
my_cov <- function(x, y) mean(x * y) - mean(x) * mean(y)
my_cov(forbes$TempC, forbes$Pressione) # = my_cov(forbes$Pressione, forbes$TempC)
# [1] 9.067404
```

- In **R** esiste anche il comando `cov` che, come nel caso della varianza, divide la sommatoria per $(n-1)$ e non n per motivi legati all'**inferenza statistica**:

```
cov(forbes$Pressione, forbes$TempC) # = 17 / 16 * my_cov(forbes$TempC, forbes$Pressione)
# [1] 9.634117
```

Domanda IV

- L'indice di **correlazione** è definito come:

$$\rho = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}}.$$

- Pertanto, possiamo calcolare la correlazione nei modo seguente:

```
my_cov(forbes$TempC, forbes$Pressione) / sqrt(my_var(forbes$TempC) * my_var(forbes$Pressione))
# [1] 0.9972227

cov(forbes$TempC, forbes$Pressione) / sqrt(var(forbes$TempC) * var(forbes$Pressione))
# [1] 0.9972227
```

- **Esercizio**. Come mai i risultati dei due comandi coincidono? Si verifichi questo fatto svolgendo **analiticamente** (carta e penna) i conti.

- In **R** esiste anche il comando `cor`, che permette di ottenere la correlazione

```
correlation <- cor(forbes$TempC, forbes$Pressione)
correlation
# [1] 0.9972227
```

Domanda V

Domanda

Si ottenga la retta ai minimi quadrati per la relazione tra `tempF` e `pressione` e la si disegni nel grafico ottenuto in precedenza.

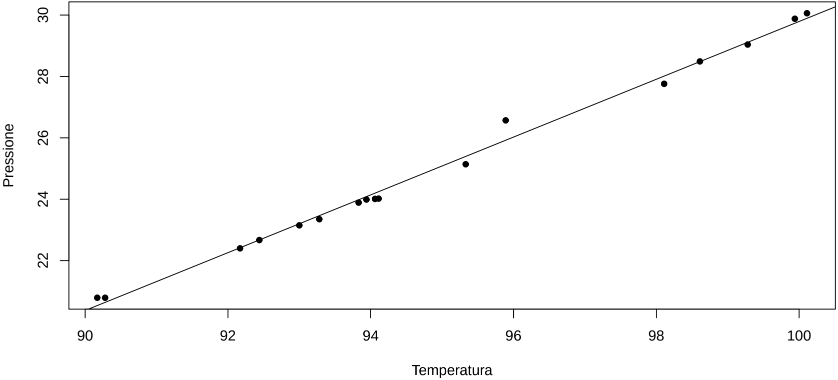
- Anzitutto ricordiamo che in un **modello lineare** del tipo $y_i = \alpha + \beta x_i + \epsilon_i$, le stime ai **minimi quadrati** sono pari a

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}, \qquad \hat{\beta} = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}.$$

- Pertanto, possiamo calcolare la correlazione nei modo seguente:

```
# Coefficiente angolare
beta_hat <- my_cov(forbes$TempC, forbes$Pressione) / my_var(forbes$TempC)
# Intercetta
alpha_hat <- mean(forbes$Pressione) - mean(forbes$TempC) * beta_hat

c(alpha_hat, beta_hat)
# [1] -64.3587103   0.9414995
```

Domanda V




```r
plot(forbes$TempC, forbes$Pressione, pch = 16, xlab = "Temperatura", ylab = "Pressione")
abline(a = alpha_hat, b = beta_hat)
```

Domanda VI

Domanda

In base al modello stimato, se la temperatura di ebollizione dell'acqua è pari 97 gradi Celsius, a quanto è pari la pressione?

- Utilizzando le stime ottenute, possiamo calcolare rapidamente i valori previsti:

```
x <- seq(from = 90, to = 100, length = 20)
alpha_hat + beta_hat * x
#  [1] 20.37625 20.87177 21.36730 21.86283 22.35835 22.85388 23.34940 23.84493
#  [9] 24.34046 24.83598 25.33151 25.82703 26.32256 26.81809 27.31361 27.80914
# [17] 28.30467 28.80019 29.29572 29.79124
```

- In particolare, quando `tempC = 97` si ha che:

```
alpha_hat + beta_hat * 97
# [1] 26.96674
```

Domanda VII

Domanda

Si ottenga un indice di bontà di adattamento ai dati della curva ottenuta e lo si interpreti nel contesto del problema.

- Il coefficiente R^2 per un modello di regressione lineare semplice è definito come:

$$R^2 = 1 - \frac{\text{var}(r)}{\text{var}(y)} = \rho^2,$$

  dove $r_1, \dots, r_n$ sono i **residui**.

- Anzitutto quindi calcoliamo i residui:

```
residuals <- forbes$Pressione - (alpha_hat + beta_hat * forbes$TempC)
```

- Il coefficiente R^2 può quindi essere ottenuto in due modi diversi:

```
correlation^2
# [1] 0.994453
1 - my_var(residuals) / my_var(forbes$Pressione)
# [1] 0.994453
```