Relazione progetto del laboratorio di Probabilità e Statistica

Introduzione

Il progetto consiste nell'analizzare i dati di vendita di un prodotto prima e dopo una campagna pubblicitaria, per stabilire se è stata efficace. Per lo svolgimento dell'analisi è stato creato uno script con il linguaggio `R`, disponibile a questo indirizzo: https://github.com/Deivmercer/Progetto-Probabilita-e-Statistica. Nel corso della relazione verranno illustrati i comandi utilizzati ai fini dell'analisi.
Come prima cosa è necessario leggere da un file i dati delle vendite; per comodità, i dati presenti nel file `Excel` sono stati riportati in un `CSV`, per poi essere letti con il seguente comando:

```
dati <- read.csv("dati.csv", sep = "\t", header = TRUE)
```

Statistica descrittiva

La prima parte dell'analisi consiste nell'utilizzare alcuni strumenti della statistica descrittiva, per avere una sintesi dei dati su cui stiamo lavorando.

```
media_prima <- mean(dati$Prima)
mediana_prima <- median(dati$Prima)
varianza_prima <- var(dati$Prima)
deviazione_standard_prima <- sd(dati$Prima)
quantili_prima <- quantile(dati$Prima, c(0.25, 0.5, 0.75))
outliers_prima <- boxplot.stats(dati$Prima)$out
```

Le istruzioni `R` sopra riportate sono state utilizzate su entrambe le rilevazioni delle vendite, ma sono state riportate una sola volta per semplicità. Usando il comando `mean` calcoliamo la media delle vendite, che è risultata essere 215.93 prima e 227.41 dopo; dunque è possibile osservare che in seguito alla campagna pubblicitaria, le vendite sono mediamente aumentate, almeno nei punti vendita presi a campione.
Con `var` otteniamo le varianze, che sono 1652.87 prima e 1868.63 dopo; la varianza equivale al quadrato della deviazione standard, che rappresenta la misura della dispersione dei dati rispetto alla media, che possiamo calcolare sia facendo la radice delle varianze (con il comando `sqrt`), sia utilizzando il comando `sd`. Otteniamo che la deviazione standard valeva 40.66 prima della campagna pubblicitaria e 43.23 dopo, e da questo è possibile dedurre che, tipicamente, le vendite rilevate prima della campagna pubblicitaria sono più vicine alla media di quelle rilevate dopo, che tipicamente hanno uno scarto maggiore. Procediamo ora a calcolare i quantili delle serie di dati, usando i comandi `quantile` (con `c` otteniamo un vettore contenente i quantili che ci interessano) che ritornano i seguenti risultati:

- Prima: primo quartile = 187.25, mediana = 216 e terzo quartile = 240.25;
- Dopo : primo quartile = 197.75, mediana = 223 e terzo quartile = 252.

I quartili (100 p-esimi percentili), rappresentano quella quantità t per cui almeno il 100p% dei dati sono <= t. Nello specifico, il primo quartile (o 25-esimo percentile) è quel valore maggiore del 25% dei dati, la mediana (o secondo quartile o ancora 50-esimo percentile) è quel valore maggiore del 50% dei dati e che quindi si trova esattamente in posizione centrale rispetto agli altri dati, mentre il terzo quartile (o 75-esimo percentile) è quel valore maggiore del 75% dei dati. Per calcolare la mediana è possibile usare anche il comando `median`.
A questo punto usiamo il comando `boxplot` per visualizzare i box plot, ovvero un grafico che mette in risalto la distribuzione dei dati.

```
boxplot(dati$Prima, dati$Dopo)
```

Di seguito sono riportati i box plot (1: prima, 2: dopo):

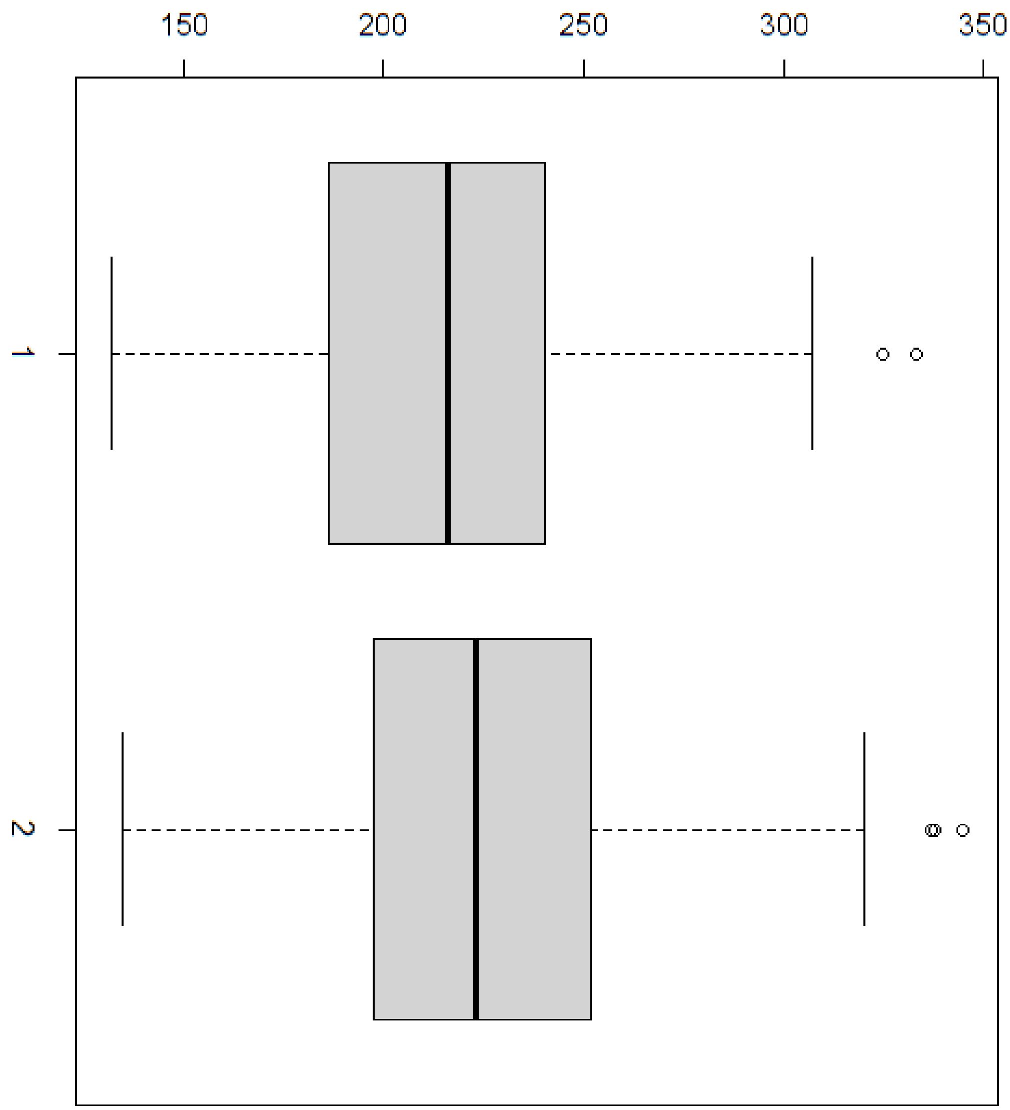

Figure 1: Box plot

Nei box plot possiamo notare che le barre più esterne a sinistra ed a destra sono rispettivamente i minimi ed i massimi, le barre a sinistra delle "scatole" sono i secondi quartili, quelle all'interno sono le mediane mentre quelle a destra sono i terzi quartili. I dati tra il minimo ed il primo quartile, tra il primo quartile e la mediana, tra la mediana ed il terzo quartile e tra il terzo quartile ed il massimo sono esattamente il 25% dei dati. I punti a destra del box plot sono dei sospetti outliers (valori anomali), cioè valori particolarmente distanti dalle altre osservazioni disponibili; utilizzando il comando `boxplot.stats(...)$out` possiamo vedere che questi sono 325 e 333 nel primo box plot e 345, 337 e 338 nel secondo. Analizzando quantili e box plot possiamo immediatamente notare come tutti i quantili del secondo box plot siano maggiori dei loro corrispettivi del primo, e che di conseguenza il box plot è posizionato di poco più a destra; questa è un ulteriore evidenza del fatto che le vendite dopo la campagna pubblicitaria sono state tipicamente più alte.

Statistica inferenziale

Effettuiamo un test T sulla differenza delle medie di due campioni normali accoppiati, per verificare che la differenza tra le medie sia diversa da 0. Per farlo, usiamo il seguente comando `R`:

```
t.test(dati$Prima, dati$Dopo, paired = TRUE)
```

Otteniamo al livello di confidenza del 95% il seguente intervallo di confidenza: (-13.59, -9.36). Poiché la statistica del test ha valore -11.48 rifiutiamo l'ipotesi nulla e accettiamo quella alternativa, perciò possiamo affermare che la differenza tra le medie è diversa da 0.
Notiamo inoltre che il p-value è 2.2e-16, che essendo molto piccolo ci permette di affermare che i dati sono in contraddizione significativa con l'ipotesi nulla.

Regressione lineare

Vogliamo ottenere la retta di regressione dei dati, per verificare se c'è una correlazione tra le vendite ottenute prima e dopo la campagna pubblicitaria. Usiamo i seguenti comandi:

```
plot(dati$Prima~dati$Dopo, data = dati)
coefficiente_di_correlazione <- cor(dati$Prima, dati$Dopo)
regressione_lineare <- lm(dati$Prima~dati$Dopo, data=dati)
abline(coef(regressione_lineare), col = "red", lwd = 5)
coefficiente_di_determinazione <- coefficiente_di_correlazione ** 2
residui <- residuals(regressione_lineare)
plot(dati$Prima~residui, data = dati)
summary(regressione_lineare)
abline(v = 0)
```

Con il primo comando plot otteniamo il grafico di dispersione dei dati, che possiamo notare siano disposti in modo non casuale:

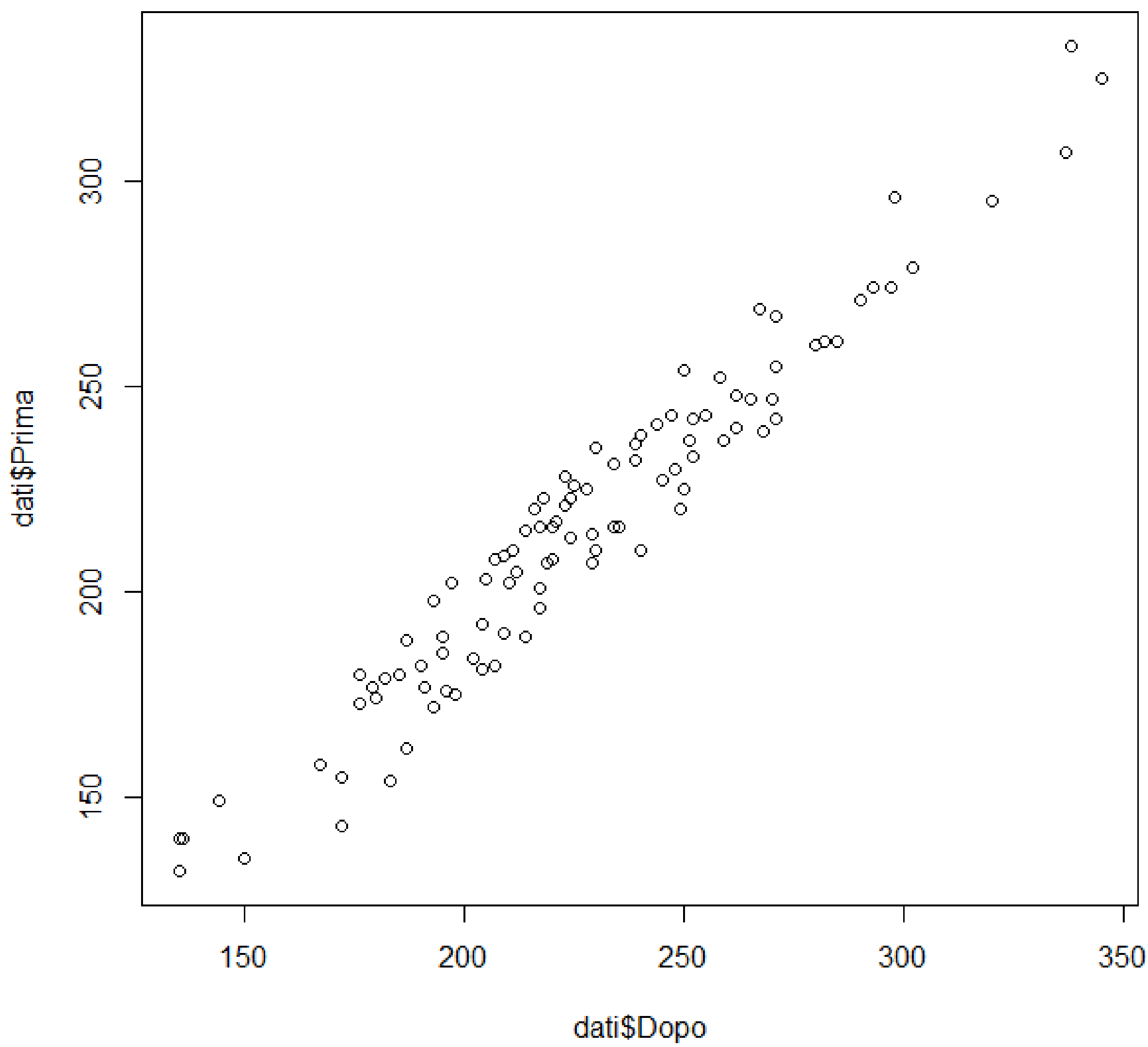

Figure 2: Grafico di dispersione dei dati

Con il comando `lm` otteniamo il modello di regressione lineare dei dati, da cui poi possiamo ottenere il coefficiente di correlazione con il comando `coef` ed i residui con il comando `residuals`. Il coefficiente di correlazione è pari a 0.97, il che significa che c'è una forte correlazione positiva tra i dati; possiamo inoltre usare questo coefficiente per disegnare sul grafico la retta di regressione, con il comando `abline`:

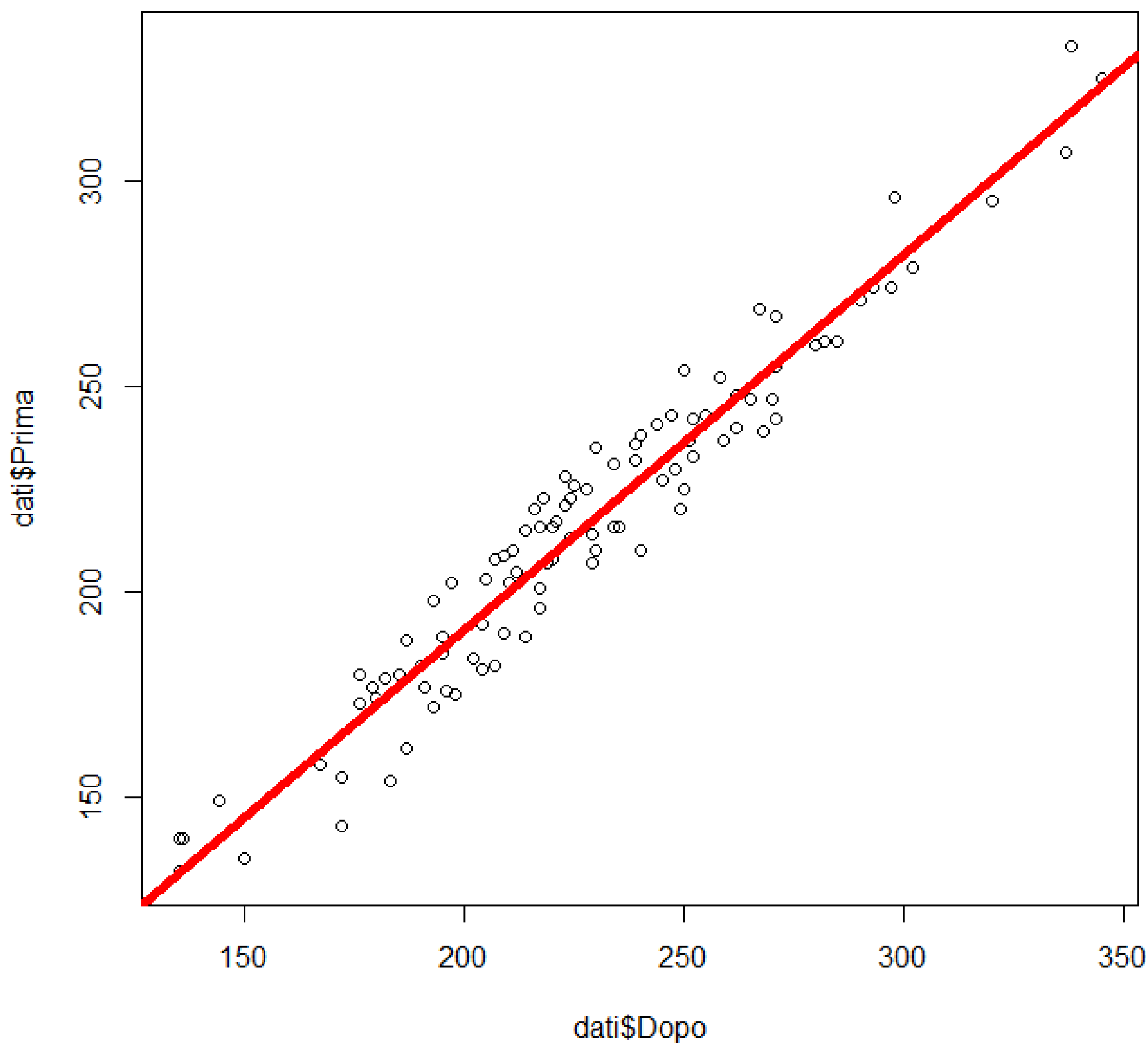

Figure 3: Grafico di dispersione con la retta di regressione

Con i residui, invece, possiamo disegnare il grafico di dispersione dei residui:

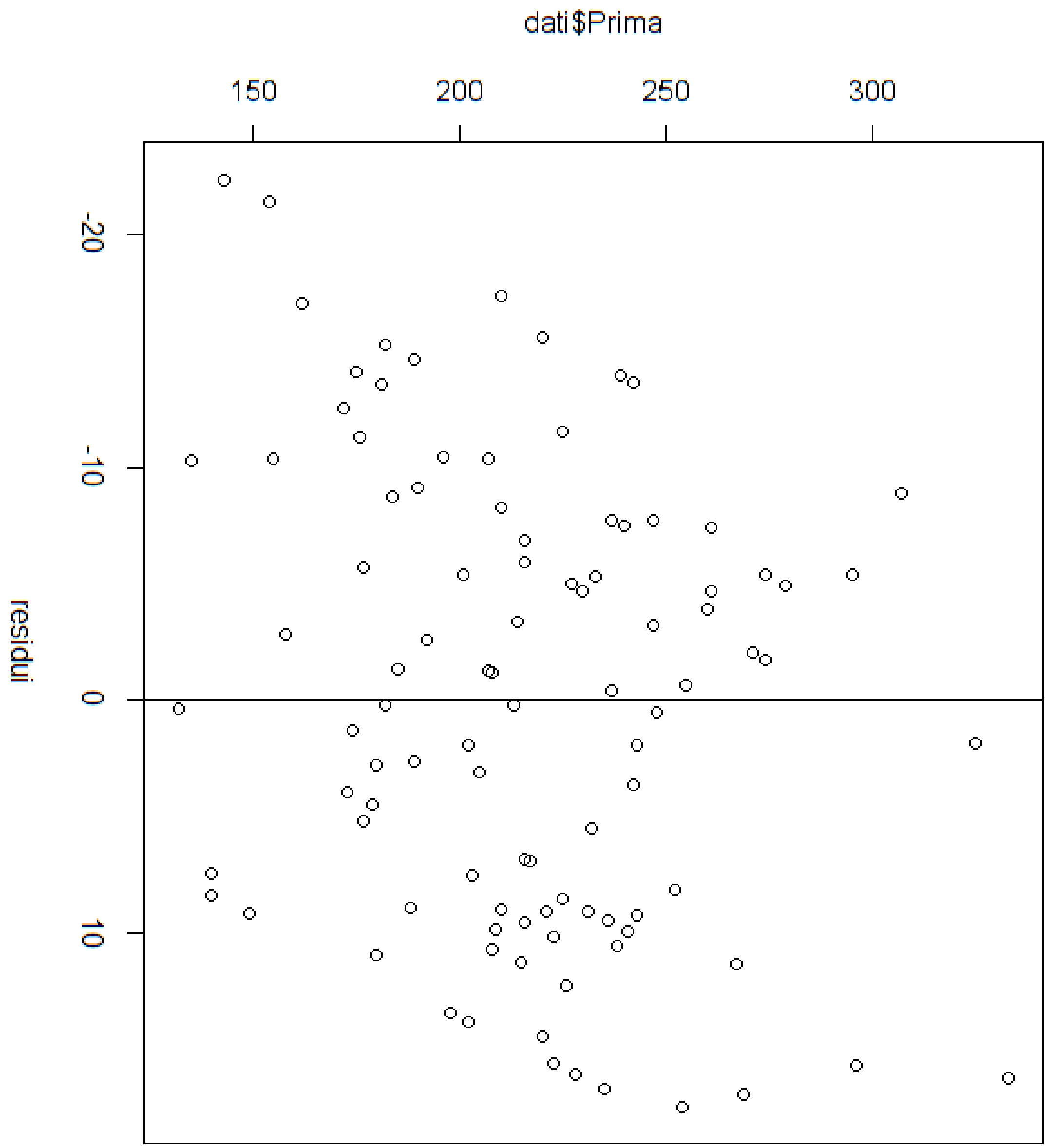

Figure 4: Grafico di dispersione dei residui

Da questo possiamo notare che i residui si dispongono in modo casuale attorno all'asse X.
Infine, usando il comando `summary` possiamo vedere che il coefficiente di determinazione è uguale a 0.94, che significa che il 94% delle risposte sono giustificate dai predittori, ed il risultato del test che verifica se β è diverso da 0.
Notare che è possibile calcolare il coefficiente di determinazione anche come il quadrato del coefficiente di correlazione.

Conclusione

In conclusione, dopo aver considerato i risultati delle analisi effettuate, possiamo affermare che in seguito alla campagna pubblicitaria c'è stato un aumento delle vendite del prodotto.