ESAME DI ANALISI STATISTICA MULTIVARIATA - Modelli Statistici — Docente Fulvia Pennoni – Statistica e Gestione delle Informazioni II ANNO

LUCINI PAIONI ANDREA 826578

30/04/2020

ESERCIZIO 1

.1

```
load("res.Rdata")
summary(res)
```

```
##        Y               X1              X2              X3        X4
##  Min.   :16.00   Min.   :15.00   Min.   : 9.00   Min.   :12.00   0:50
##  1st Qu.:34.00   1st Qu.:19.00   1st Qu.:16.00   1st Qu.:18.00   1:50
##  Median :41.00   Median :21.00   Median :18.00   Median :19.00
##  Mean   :41.95   Mean   :21.18   Mean   :17.93   Mean   :19.23
##  3rd Qu.:48.00   3rd Qu.:23.00   3rd Qu.:20.00   3rd Qu.:20.25
##  Max.   :74.00   Max.   :28.00   Max.   :26.00   Max.   :27.00
```

Abbiamo un dataset con cinque variabili, di cui: la variabile risposta Y che indica il costo, e le quattro variabili esplicative X1 (qualità del cibo), X2 (eleganza del locale), X3 (qualità del servizio) e X4, variabile binaria che indica l'ubicazione del locale, se in città o fuori città

Dalle statistiche descrittive ottenute osserviamo che la variabile risposta Y si distribuisce in un intervallo che va da un minimo di 16 e un massimo di 74; inoltre osserviamo che presumibilmente la variabile ha distribuzione simmetrica, in quanto la mediana e la media sono quasi coincidenti e la distanza del primo quartile dalla mediana è uguale alla distanza tra mediana e terzo quartile. Tuttavia la distanza tra terzo quartile e valore massimo (di 26 euro) è più ampia di quella tra valore minimo e primo quartile (di soli 18 euro), quindi possiamo immaginare una coda della distribuzione più "dispersa" nei valori elevati di costo.

Per quanto riguarda le variabili esplicative, osserviamo che X1, X2 e X3 hanno tutte un campo di variazione meno ampio di quello della variabile risposta Y, e che in tutte e tre le distribuzioni la media e la mediana sono pressochè coincidenti. Sono dunque distribuzioni abbastanza simmetriche, con leggere differenze solo nelle code (ad esempio tra minimo e primo quartile e terzo quartile e massimo in X1).

L'ultima variabile esplicativa X4, è binaria, e può assumere solo valore 0 o 1, valori equamente distribuiti (50 per livello sulle 100 osservazioni totali).

Proviamo ora a visualizzare le prime osservazioni, per vedere se già da queste possiamo ricavare qualche informazione su eventuali correlazioni tra le variabili.

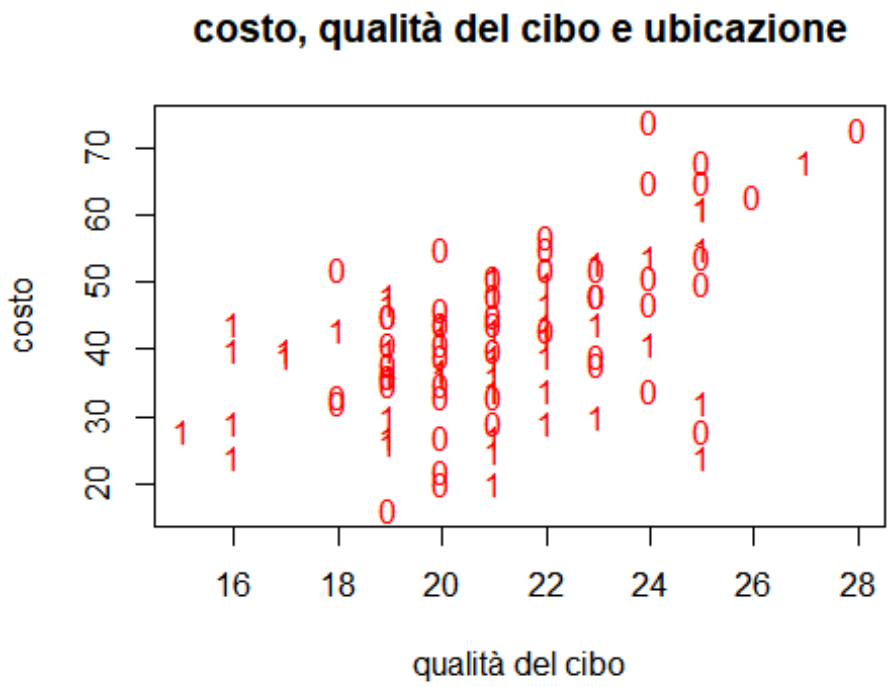
```
head(res)
```

```
##    Y X1 X2 X3 X4
## 1 65 24 21 24  0
## 2 48 23 20 20  0
## 3 32 18 18 17  0
## 4 27 20 18 20  0
## 5 47 24 20 23  0
## 6 45 19 17 19  0
```

Dalle prime 6 osservazioni, notiamo che a valori elevati di costo (come nella prima osservazione), corrispondono i valori più elevati di X1, X2 e X3. Allo stesso tempo però, al valore più basso di costo Y (osservazione 4) non corrispondono i valori più bassi delle tre variabili. Tuttavia c'è da tenere conto che si tratta solo delle prime 6 osservazioni, per avere un'idea dell'andamento generale delle variabili conviene utilizzare grafici e indici che definiscano la correlazione tra le variabili.

Interessante, inoltre, è la sola presenza di valore 0 per la variabile X4 nelle prime 6 variabili osservate nel dataset.

```
plot(Y ~ X1, pch=as.character(X4),res, col="red",main="costo, qualità del cibo e
ubicazione",xlab="qualità del cibo",ylab="costo")
```



Dal seguente grafico osserviamo che a qualità del cibo scadente corrispondono in generale ristoranti ubicati in città con costi dei pasti abbastanza contenuti, mentre notiamo che tra i ristoranti con la migliore qualità del cibo ci sono soprattutto ristoranti fuori città, ma si tratta tuttavia dei ristoranti più costosi (solo un ristorante di città rientra tra questo gruppo).

.2

```
library(faraway)
vif(res[1:4])
```

```
##        Y       X1       X2       X3
## 3.099810 2.163548 2.279656 2.802223
```

La quota di inflazione della varianza (in inglese VIF) indica la presenza o assenza di collinearità eccessiva tra le variabili presenti in un dataset. Ovviamente, si ricercano sempre valori di collinearità bassi (quindi intorno a 1), che ci garantiscono che gli errori standard dei parametri non siano condizionati da associazione tra le variabili esplicative presenti nel dataset.

In questo caso, i valori sono abbastanza bassi (in quanto consideriamo problematici valori superiori a 5).

```
sqrt(3.099)
```

```
## [1] 1.760398
```

Ad esempio, l'errore standard stimato del costo è circa 1.5 volte maggiore di quanto sarebbe stato se le variabili esplicative non fossero state assolutamente associate tra di loro.

.3

```
lm1<-lm(Y ~ X1 + X2 +X3 +X4, res)
summary(lm1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = res)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.2214  -3.7018   0.5539   5.0533  14.9392
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -32.6075     6.3409  -5.142 1.45e-06 ***
## X1            0.9453     0.3790   2.494  0.01435 *
## X2            1.8051     0.2351   7.679 1.42e-11 ***
## X3            1.2065     0.4235   2.848  0.00539 **
## X41          -2.0584     1.3786  -1.493  0.13873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.808 on 95 degrees of freedom
## Multiple R-squared:  0.6848, Adjusted R-squared:  0.6715
## F-statistic:  51.6 on 4 and 95 DF,  p-value: < 2.2e-16
```

Il modello di regressione stimato ha un buon valore di R^2, per cui ci aspettiamo che "spieghi" abbastanza bene i valori di Y

La terza colonna della tabella dei coefficienti indica il valore del t value, mentre la quarta colonna indica il p-value corrispondente al test. Si tratta di due valori importanti che ci permettono di verificare l'ipotesi nulla H0 per cui il parametro corrispondente viene posto nullo (ovvero uguale a 0) nel modello di regressione. Concentrandoci sulla variabile X4, osserviamo che il valore del p-value è pari a 0.13, dunque non esiste evidenza contro l'ipotesi nulla di escludere la variabile, dunque possiamo decidere di escludere la variabile X4 dal modello di regressione.

.4

```
lm2<-lm(Y ~ X1 + X2 +X3, res)
summary(lm2)

##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = res)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2880  -4.2114   0.2809   4.6024  15.7294
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -35.0757     6.1607  -5.693 1.35e-07 ***
## X1            0.9956     0.3799   2.621   0.0102 *
## X2            1.8186     0.2364   7.692 1.26e-11 ***
## X3            1.2133     0.4262   2.847   0.0054 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.851 on 96 degrees of freedom
## Multiple R-squared:  0.6774, Adjusted R-squared:  0.6673
## F-statistic: 67.19 on 3 and 96 DF,  p-value: < 2.2e-16
```

Abbiamo stimato ora il modello di regressione escludendo la variabile esplicativa X4 (riferita all'ubicazione del ristorante). Dal summary, otteniamo alcune indicazioni riguardanti: - i residui (nella tabella "residuals"), che vanno da un minimo di -21 circa a un massimo di 15 (osserviamo che la mediana è leggermente sfasata rispetto a 0, e che anche i valori massimo e minimo hanno diversa distanza dal valore 0). - i coefficienti del modello di regressione (prima colonna della tabella coefficients), che ci mostrano un valore di -35 per l'intercetta (ovvero la variabile risposta assume valore 35 se tutte le altre variabili esplicative vengono poste uguali a 0), e tutti valori positivi per le tre variabili esplicative. Facendo un esempio, a parità di valore per le variabili X2 e X3, la variabile risposta Y aumenta di circa un'unità in caso di aumento unitario della variabile X1. Nella tabella abbiamo, inoltre, anche gli errori standard per ogni variabile, e i t value e p-value dei coefficienti. - altre statistiche, come l'R^2, l'indice di determinazione multipla che ci indica quanta variabilità di Y è "spiegata" dal modello di regressione. O anche

Le proprietà degli stimatori del modello di regressione sono:

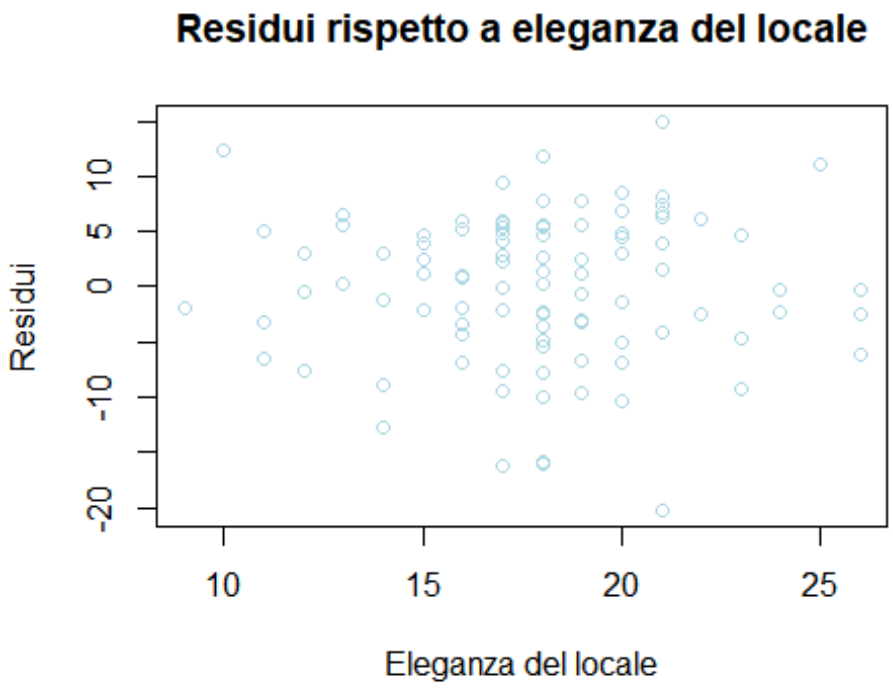
1. la somma dei residui nel modello di regressione che comprende l'intercetta è uguale a 0. Quindi la somma dei valori interpolati è uguale alla somma dei valori osservati.

2. L'ortogonalità dei residui di rispetto ad ogni variabile esplicativa.

3. Le variabili Z e Y (Y cappello) sono incorrelate, quindi la cov(Z,Y)=0

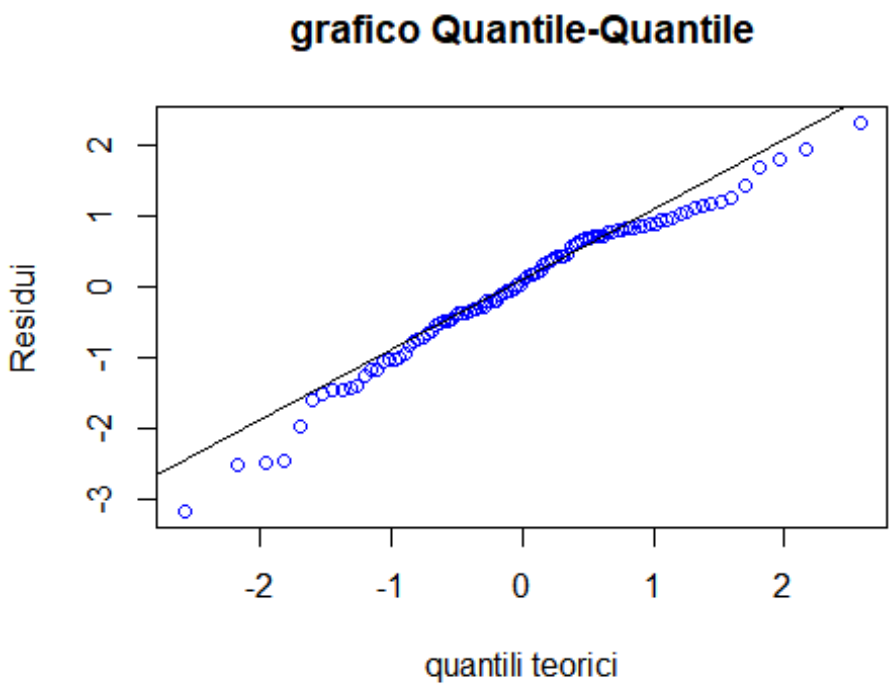
.5

```
plot(res$X2,lm1$residuals,main="Residui rispetto a eleganza del
locale",xlab="Eleganza del locale", ylab="Residui", type="p",col="light blue")
```



Dal grafico che mette in relazione i residui rispetto all'eleganza del locale, notiamo una forma simile ad un'ellisse, con alcuni punti che si discostano leggermente da questa forma (un paio di punti con alti residui e alto livello di eleganza, un punto con alti residui e basso livello di eleganza). Ma in generale, non si osservano pattern inaspettati.

```
rT<-rstudent(lm1)
qqnorm(rT,main="grafico Quantile-Quantile",xlab="quantili
teorici",ylab="Residui",col="blue")
qqline(rT)
```



Dal grafico Quantile-Quantile, invece, osserviamo che i dati seguono abbastanza l'andamento della retta nella parte centrale della distribuzione, mentre abbiamo degli scostamenti dalla stessa per i valori più bassi e per i valori più alti dei residui. In particolare, due punti sono tra i più distanti dalla retta, rispettivamente il valore più basso e il valore più alto dei residui.

.6

Gli intervalli congiunti per coppie di coefficienti di regressione si costruiscono con delle ellissi, che ci consentono di valutare l'ipotesi nulla (congiunta) di nullità dei coefficienti

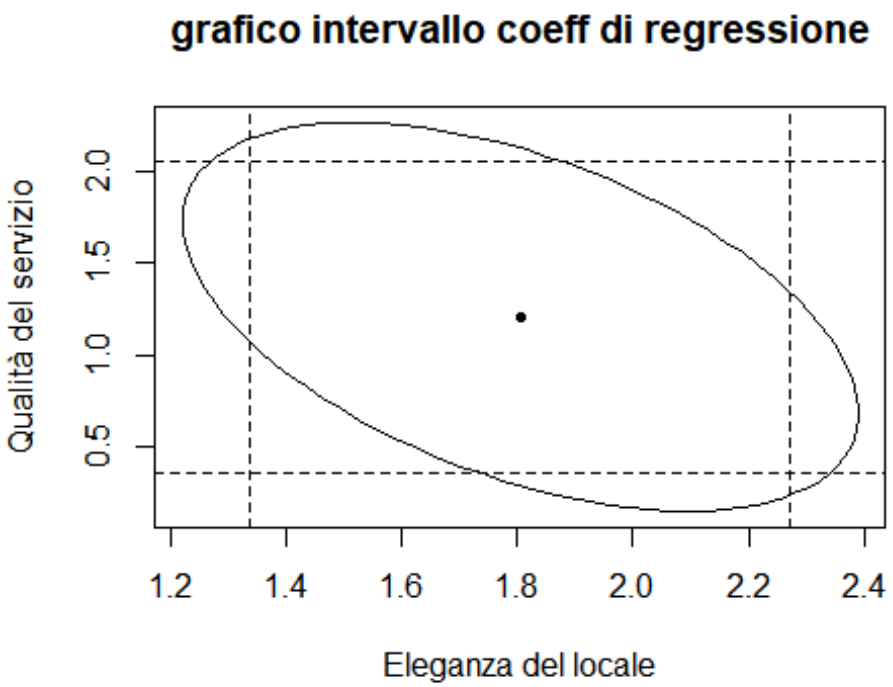
```
require(ellipse)

## Loading required package: ellipse

##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs

plot(ellipse(lm1,c(3,4)), type="l",xlab="Eleganza del locale",ylab="Qualità del
servizio",main="grafico intervallo coeff di regressione")
points(coef(lm1)[3],coef(lm1)[4],pch=20)
abline(v=confint(lm1)[3,],lty=2)
abline(h=confint(lm1)[4,],lty=2)
```



Dal grafico ottenuto, osserviamo che l'ellisse è orientata verso sinistra, questo indica una correlazione negativa tra i due coefficienti stimati. Inoltre, osserviamo il punto al centro dell'ellisse, che sono i valori della stima puntuale dei due coefficienti, e i quattro assi (due orizzontali e due verticali), che indicano gli intervalli di confidenza dei singoli coefficienti.

Possiamo, inoltre, rifiutare il test che valuta l'ipotesi congiunta di nullità dei due coefficienti, in quanto l'origine (0,0) non è compresa all'interno dell'ellisse.

.7

```
res2<-res[,-5]
names(res2)

## [1] "Y"  "X1" "X2" "X3"

xc<-c(1,23,19,27)
names(xc)<-c("Intercept","X1","X2","X3")
predict(lm2,new=data.frame(t(xc)),interval="confidence")

##        fit      lwr      upr
## 1 55.13495 49.49898 60.77092
```

In caso di previsione del valore medio, l'intervallo di previsione è più piccolo rispetto all'intervallo di previsione della risposta (infatti in questo caso l'intervallo va da 50 circa a 60 circa). Il valore puntuale è indicato da fit, e assume un valore di 55.1 circa.

.8

Il criterio di Akaike permette di selezionare il modello migliore, considerando o no alcune delle variabili in studio (oltre all'intercetta). Si può utilizzare il procedimento "forward" (parte dal modello nullo e aggiunge una variabile per volta), "backward" (parte dal modello completo e toglie una variabile per volta) o "stepwise" (un misto dei due precedenti, per cui vengono considerate o tolte variabili dal modello fino a quando non si trova il miglior modello possibile). Si ricerca sempre il valore minimo di AIC, per cui verrà indicato il modello migliore.

```
step(lm1, direction="forward")
```

```
## Start:  AIC=388.49
## Y ~ X1 + X2 + X3 + X4

##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = res)
##
## Coefficients:
## (Intercept)           X1           X2           X3          X41
##    -32.6075       0.9453       1.8051       1.2065      -2.0584
```

In questo caso, il valore minimo di AIC è di 388.49, quello per cui il modello è completo, quindi è consigliabile mantenere in analisi il modello completo.