Esercizi - Analisi Predittiva

Modelli lineari semplici e multipli

Esercizio 1

I dati in `grain.dat` sono stati raccolti nel 2007 in uno studio sula relazione tra la resa in termini di alcool nel processo di distillazione e l'azoto contenuto nel grano distillato. I dati sono stati raccolti in quattro diverse aree del Regno Unito. Il dataset ha tre colonne: `nitrogen` è la percentuale di azoto (per kilogrammo), `alcohol` è la resa in alcool in Litri per Tonnellata, e`location` indica il luogo in cui è stato coltivato il grano. [Il dataset è stato reso disponibile da Julian Faraway.]

La relazione tra la resa in termini di alcool e l'azoto contenuto nel grano può essere indagata con il seguente modello lineare:

$$\text{alcohol}_i = \alpha + \beta\ \text{nitrogen}_i + \epsilon_i \tag{1}$$

1. Si produca un grafico dei dati. La relazione tra le variabili in esame appare lineare?
2. Si dia una stima puntuale per α e β.
3. Si dia una stima intervallare ad un livello di confidenza di 99% per α e β.
4. Quali sono le assunzioni necessarie per poter aver stime puntuali per i valori α e β? Quali sono le assunzioni necessarie per poter ottenere delle stime intervallari per α e β?
5. Si aggiunga la retta delle relazione stimata tra `alcohol` e `nitrogen` al grafico ottenuto al punto 1.
6. Il dataset contiene la variabile `location`. Si scriva in forma estesa il modello che R stima quando si usa la funzione `lm(alcohol location, data = grain)`.
7. É valida l'affermazione che la variabile `location` spiega una buona parte della variabilità della variabile `alcohol`?
8. Se si aggiunge la variabile `location` al modello in eq. (1) in cui solo `nitrogen` era presente nel modello, l'aggiunta di `location` risulta significativa? Come si può misurare l'evidenza contro la non-inclusione di `location` nel modello?
9. Si produca un grafico della relazione tra `location` e `nitrogen` - cosa si può notare?
10. Come si spiega la differenza dei p-value per `location` nei modelli stimati al punto 6 e al punto 8?
11. Usando il modello specificato in eq. (1): si predica il valore medio della resa di alcool per del grano contenente il 1.9% e il 2.7% di azoto per kilogrammo.
12. Si stimino gli intervalli di confidenza al 95% per i valori medi della resa di alcool stimati al punto 11. Quale è l'ampiezza di questi intervalli: si spieghi la differenza nell'ampiezza.

13. Usando il modello specificato in eq. (1): si predica il valore effettivo della resa di alcool per del grano contenente il 1.9% e il 2.7% di azoto per kilogrammo. Si dia una anche una valutazione degli intervalli predittivi al 95% per questi valori.

Esercizio 2

I dati nel file `hotel.csv` contengono informazioni sulla temperatura (`X`) e sul livello di occupazione di tre hotel (`Y`) in tre diverse città.

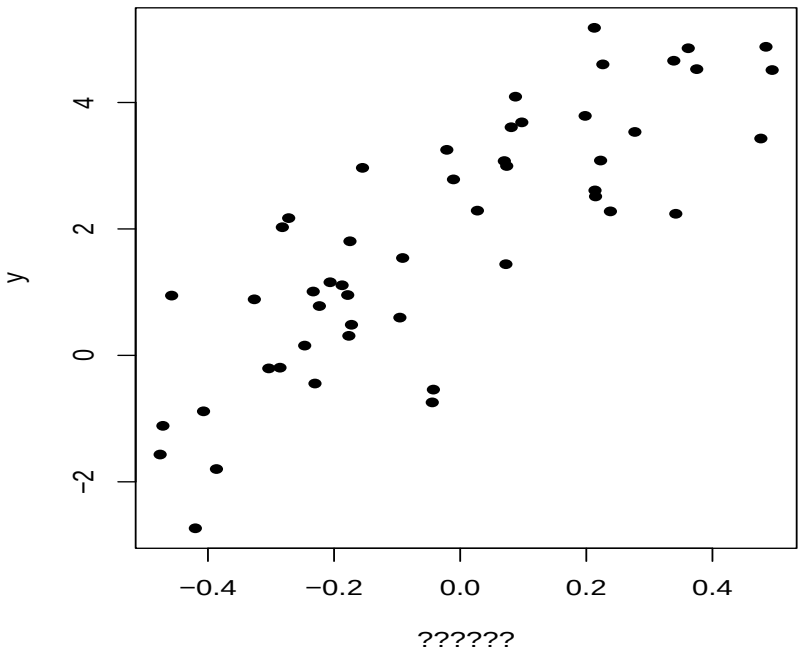
1. Si producano grafici di dispersione (o altri grafici che si ritengono utili) per valutare la relazione tra le variabili presenti nel dataset

2. Si stimi un modello lineare in cui si assume che il tasso di occupazione degli alberghi dipenda dalla temperatura: si dia un'interpretazione del modello stimato

3. Si stimino separatamente tre modelli lineari per ogni città per studiare come il tasso di occupazione degli alberghi dipende dalla temperatura. Si dia un'interpretazione dei tre modelli confrontando i risultati con quelli ottenuti al punto 2.

4. Alla luce dei modelli stimati al punto 3 - si specifichi un modello che si ritiene possa essere utile per spiegare nella maniera migliore possibile la variabilità dell'occupazione degli hotel al variare della temperatura in tutte le città incluse nel dataset.

Esercizio 3



1. Il grafico qui sopra mostra la relazione tra la variabile X e Y di interesse. Qui sotto vengono riportati i `summary` di due modelli stimati: uno usando la X mostrata in figura e uno usando un'altra variabile. Si identifichi il summary che corrisponde alla relazione mostrata in figura.

S1

```
> summary(lm(y ~ x1))

Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49684 -0.68150  0.03744  0.78701  1.88648

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0185     0.1595   12.65  < 2e-16 ***
x1            5.9989     0.5858   10.24 1.16e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.122 on 48 degrees of freedom
Multiple R-squared:  0.686,        Adjusted R-squared:  0.6795
F-statistic: 104.9 on 1 and 48 DF,  p-value: 1.157e-13
```

S2

```
> summary(lm(y ~ x2))

Call:
lm(formula = y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9848 -1.6183  0.4791  1.2761  3.7511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.50848    1.08136   3.245  0.00215 **
x2          -0.03306    0.02086  -1.585  0.11960
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.952 on 48 degrees of freedom
Multiple R-squared:  0.04972,        Adjusted R-squared:  0.02992
F-statistic: 2.511 on 1 and 48 DF,  p-value: 0.1196
```

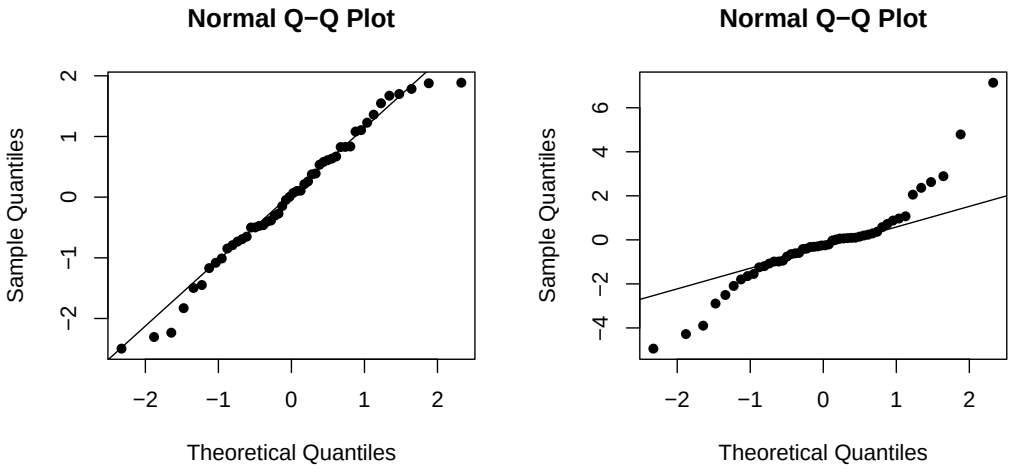
2. Come si può interpretare il seguente output di R?

```
> anova(lm(y ~ x1), lm(y ~ x1+x2))

Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 60.451
2     47 58.730  1    1.7209 1.3772 0.2465
```

3. Qui sotto vengono mostrati tre valori di R^2 e i tre modelli da cui sono stati estratti: si accoppino i modelli e valori di R^2 ad essi corrispondenti:

Modello	R^2
`lm(y ~ x1+x2)`	0.686
`lm(y ~ x1)`	0.0497
`lm(y ~ x2)`	0.695

4. Quale dei grafici quantile-quantile indica un comportamento del campione analizzato più simile alla distribuzione di riferimento? Si spieghi come i grafici quantile-quantile possono essere utilizzati quando si stimano modelli lineari.



Esercizio 4

Si prenda in esame il dataset `prostate` dal pacchetto R `faraway`. Si desidera stimare la relazione tra la un certo antigene (descritto dalla variabile `lpsa`) e altre variabili contenute nel dataset.

1. Si prenda in considerazione un modello lineare multiplo in cui tutte le variabili presenti nel dataset sono usate come predittori (Modello 1). Si stimi il modello e usando la funzione `summary` (o equivalenti) si trovi il valore della statistica F della significativitá globale del modello. Si interpreti il valore della statistica test.

2. Si trovi la stima puntuale del coefficiente di regressione legato alla variabile `age` dentro al modello 1: che interpretazione si può dare al valore del coefficiente? Si produca un grafico di dispersione (scatter plot) della variabile `age` e la variabile `lpsa`: come si può interpretare il coefficiente di regressione identificato per il modello 1 alla luce del grafico?

3. Si trovi l'intervallo di confidenza al 90% e 99% per il coefficiente di regressione legato alla variabile `age` dentro al modello 1: che interpretazione si può dare ai due intervalli? Cosa si può dedurre da questi intervalli di confidenza sul p-value della variable `age` nel Modello 1?

4. Si stimino intervalli di confidenza per il valore di `lpsa` di due pazienti con le seguenti caratteristiche:

```
> nd <- prostate[1,-9] ## to ensure correct names
> nd[1,] <- c(1.45, 3.62,65,0.3,0,-0.8,7,15)
> nd[2,] <- c(4,6.2,83,2.33,1,2.96,9,100)
> rownames(nd) <- c("Patient A", "Patient B")
> nd
```

```
          lcavol lweight age lbph svi   lcp gleason pgg45
Patient A   1.45    3.62  65 0.30   0 -0.80       7    15
Patient B   4.00    6.20  83 2.33   1  2.96       9   100
```

   Si commenti l'ampiezza degli intervalli di confidenza.

5. Si stimi ora un nuovo modello, modello 2, in cui solo le variabili predittive con un p-value del test di significatività nel Modello 1 minore di 0.05.

6. Si usi questo nuovo modello per stimare i valori dei pazienti A e B: si commenti sull'ampiezza degli intervalli di confidenza trovati nel Modello 1 e Modello 2.

7. Si testi al significatività del modello 2 contro il modello 1, esplicitando l'ipotesi nulla e alternativa sotto studio.

   Per caricare il dataset nella propria workspace è necessario avere il pacchetto faraway installato - questo si può fare una volta sola con il comando install.packages("faraway"). Successivamente sarà necessario usare il comando data(prostate, package = "faraway").

Esercizio 5

Si prenda in considerazione il dataset 'ex5' che si può costruire seguendo i dettagli del file ex5.R su Moodle.

La variabile X descrive il numero di ore che settimanalmente degli studenti investono nello studio di una materia. La variabile Y è una misure di performance standardizzata per l'anno scolastico degli studenti nella materia: valori più alti indicano performance migliori. La variabile Z indica l'anno in cui ogni studente è iscritto.

1. Si stimi la relazione tra X e Y: come si può interpretare il valore stimato del coefficiente angolare?

2. Si proceda ad una verifica di ipotesi della significatività del coefficiente angolare: si specifichino ipotesi nulla ed alternativa

3. Si proceda ad una verifica di ipotesi sul coefficiente angolare β_1 in cui l'ipotesi nulla è $H_0 : \beta_1 < 0$. Qual è la ipotesi alternativa? Come cambia il p-value rispetto al p-value calcolato per il punto (ii)?

4. Si stimi ora un modello lineare multiplo in cui sia X che Z sono inclusi come predittori: qual è ora la stima del coefficiente angolare legato ad X? Si commenti su cosa può essere la causa del cambiamento della stima

5. Si produca una rappresentazione grafica che esemplifichi i valori stimati per $E[Y]$ ottenuti nei due modelli di regressione stimati fino ad ora

6. E' corretto usare la variabile Z come variabile numerica? Se no, che cambiamento si potrebbe apportare in R alla variabile? Che implicazioni ha questo per la forma del modello stimato (per esempio sul numero di parametri stimato)?

Esercizio 6

Si prenda in esame il dataset `Davis` dal pacchetto R `carData`. Si desidera stimare la relazione tra il peso dichiarato (`repwt`) da uomini e donne e il peso misurato (`weight`) da uomini e donne.

1. Si stimi tre modelli di crescente complessità in cui la relazione è stimata essere la stessa per entrambi i sessi, viene permesso all'intercetta di essere diversa per i due sessi e infine in cui si permette a intercetta e coefficiente angolare di essere diversi per i due sessi.

2. Si scrivano in forma estesa (in formula matematica) i tre modelli specificati al punto 1 e si confronti la bontà di adattamento dei tre modelli indicando quale modello viene scelto come modello finale.

3. Si verifichi se per il modello selezionato valgono le assunzioni alla base della costruzione dei modelli lineari. Si commenti in particolare se sono presenti punti particolarmente influenti sulla stima.

Esercizio 7

[Esercizio di Esame aa 2019/2020 - prof. Gaetan]

Si consideri il modello di regressione

$$Y_i = \begin{cases} \beta_1 + \varepsilon_i & i = 1, \dots, 5 \\ \beta_1 + \beta_2(i-5) + \varepsilon_i & i = 6, \dots, 10 \end{cases}$$

dove ε_i, $i = 1, \dots, 10$ sono v.c. casuali $\mathcal{N}(0, \sigma^2)$ indipendenti.

1. Si specifichi se le assunzioni usualmente adottate in un modello di regressione lineare Gaussiano (linearità della relazione, normalità ed omoschedasticità degli errori, indipendenza delle osservazioni) sono soddisfatte dal modello sopra riportato.

2. Il modello può essere scritto nella forma matriciale $Y = X\beta + \varepsilon$. Si dia l'espressione della matrice X.

3. Si argomenti quale possa essere la distribuzione dello stimatore di massima verosimiglianza per β.

4. Supponendo di aver ottenuto le stime $(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}) = (2.86, 0.1, 0.86)$, si derivino gli intervalli di confidenza con livello esatto 0.90 per β_1 e β_2.

5. Di quale altra informazione ci sarebbe bisogno per calcolare R^2 ?