Esercitazione 2 per l'analisi dei dati

Esercizio 1

a)

Avendo a disposizione i dati di 565 persone per quanto concerne l'età (in anni), altezza (in cm), peso (in kg), suddivisi per genere 1 e genere 2, per individuare la codifica utilizzata si procede suddividendo i suddetti dati per classi di età in rapporto alle altezze. Non essendo specificata la provenienza delle persone analizzate, prenderemo come riferimento l'altezza media delle donne e degli uomini a livello mondiale ovvero 159,5 cm e 171 cm. Si è deciso inoltre di non tener conto della fascia di età tra i 16 e 18 anni in quanto ogni ragazzo/ragazza raggiunge l'altezza definitiva intorno ai 18 anni.

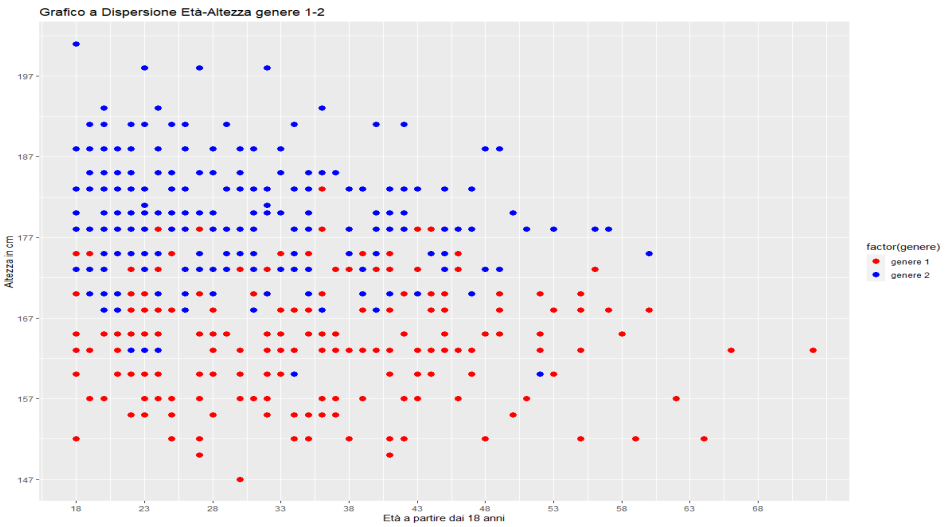


Figura 1

Nella Figura 1, il grafico a dispersione indica che le altezze del genere 2 sono maggiori in buona parte rispetto al genere 1. Si presume quindi che il genere 2 sia riferito al genere maschile e il genere 1 al genere femminile.

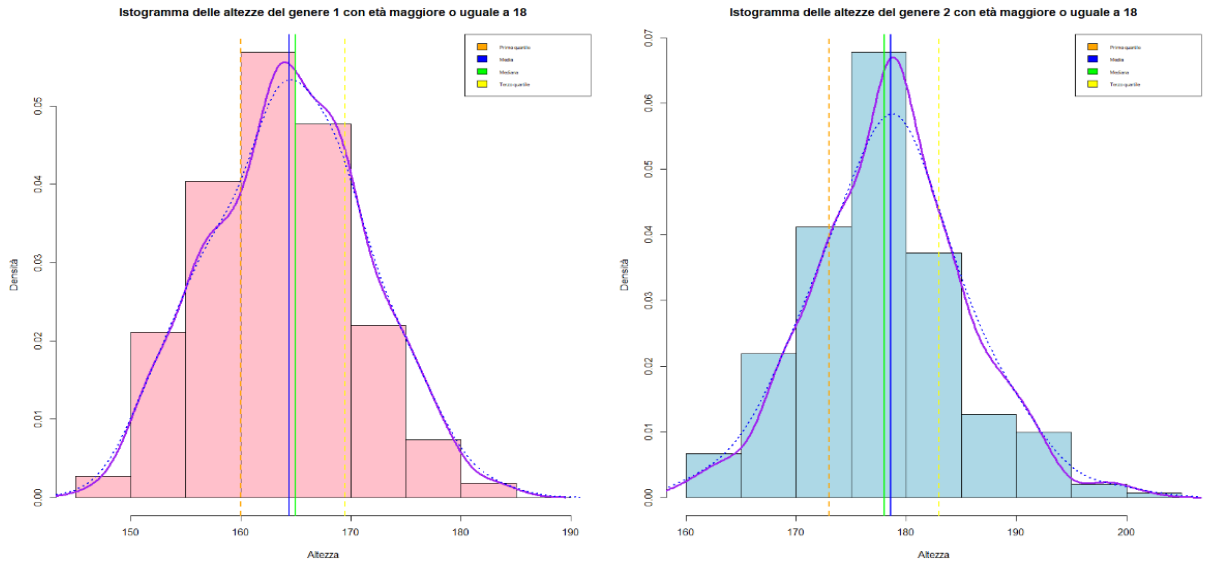
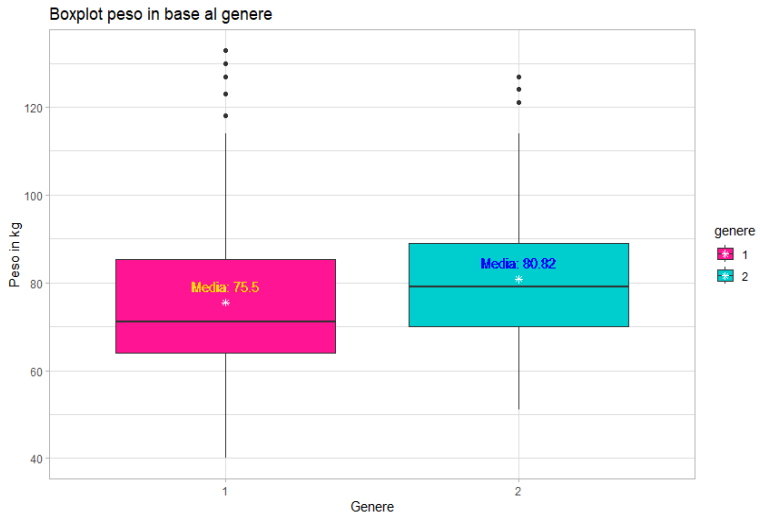


Figura 2,3

Nella Figura 2, l'altezza del genere 1, che si presuppone essere il genere femminile, varia dai 147 cm a 183 cm con una media di 164,42 cm e una deviazione standard di 7. Il 72% dei valori delle altezze si trova negli intervalli che vanno da 155-170 cm con pochi casi estremi sia verso il basso che verso l'alto; quindi, le altezze del genere 1 sono

regolari intorno alla media. Nella Figura 3 l'altezza del genere 2, che si presuppone essere il genere maschile, varia dai 160 cm a 201 cm con una media di 168 cm e una deviazione standard di 6,88. La distribuzione è unimodale simmetrica e il 33% dei valori si trova nell'intervallo 175-180 cm. Rispetto al genere 1 ci sono più casi estremi di persone molto basse o molto alte. Si può notare che la curva di densità risulta più appuntita della curva di Kernel (tratteggiata in blu) nel picco di densità

b)



	PESO GENERE 1	PESO GENERE 2
Numero di osservazioni	236	329
Media	75.5	80.82
Deviazione standard	16.42	14.29
Minimo	40	51
Massimo	133	127
Primo quartile	64	70
Mediana (secondo quartile)	71	79
Terzo quartile	85	89
Simmetria/ Asimmetria	Asimmetria positiva in quanto (Q3 - Q2) > (Q2 - Q1)	Simmetrica
Outliers	118,123,127, 130,133	121,124,127,127,127

Tabella 1

Figura 4 (sono inclusi i pesi di tutte le età)

Con pesi e altezze possiamo calcolare l'IMC (indice di massa corporea) di entrambi i generi, classificandoli in normopeso, sovrappeso e sottopeso. Bisogna poi tenere conto che mentre l'obesità negli adulti viene misurata rapportando il peso all'altezza, lo stesso rapporto non può essere applicato ai bambini che nelle varie fasce di età hanno peso e altezza variabile, pertanto, sono stati presi in considerazioni i dati con le fasce di età maggiori o uguali di 18.

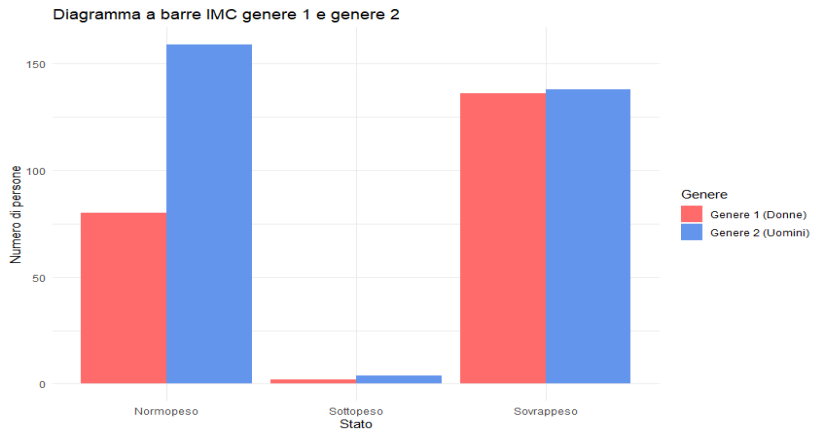


Figura 5

Nella Figura 5, notiamo una prevalenza di genere 2 (Uomini) in normopeso (IMC 18,50 - 24,99) rispetto al genere 1 (Donne).
Il 53% degli uomini è in uno stato di normopeso. In contrapposizione però il 62% del genere 1 (Donne) è in una condizione di sovrappeso (IMC 25,00-29,99) forse a causa di fattori ormonali. Per quanto riguarda lo stato di sottopeso (IMC 16,00 - 18,49) non risultato valori significativi.

c)

Regression Equation

Peso= -54,3 + 0,7633 altezza - 3,99 genere + 0,2420 età

Coefficienti

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-54,3	13,5	-4,02	0,000	
Altezza	0,7633	0,0852	8,96	0,000	1,96
Genere	-3,99	1,67	-2,39	0,017	1,92
Età	0,2420	0,0571	4,24	0,000	1,07

Il modello di regressione lineare multipla utilizza le variabili indipendenti altezza, genere ed età per prevedere il valore della variabile dipendente peso. Il valore di -54,3 rappresenta il peso previsto quando tutte le altre variabili sono uguali a zero. Il coefficiente associato alla variabile altezza indica che un aumento di 1 cm in altezza è associato ad un aumento di 0,7633 kg nel peso previsto (-54,3), mantenendo costanti le altre variabili. Il coefficiente di -3,99 associato alla variabile genere indica che, mantenendo costanti le altre variabili, gli individui di genere 2 hanno un peso previsto mediamente minore di 3,99 kg rispetto a quelli di genere 1.
Il coefficiente associato alla variabile età indica che, mantenendo costanti altezza e genere, un aumento di 1 anno di età è associato ad un aumento di 0,2420 kg nel peso previsto.
Per ogni variabile indipendente il valore del P-value con un $\alpha = 0,1$ è significativo.
Dal modello si evidenzia che la variabile indipendente del genere è quella che incide di più sulla variabile dipendente peso. Tutte le variabili hanno valori VIF inferiori a 2 e questo indica che la correlazione tra le variabili indipendenti è relativamente bassa e che il modello è stabile

S	R-sq	R-sq(adj)	R-sq(pred)
14,1625	16,16%	15,72%	14,93%

Il valore del R^2 aggiustato (R-sq(adj)) risulta molto basso per questo modello. Il valore è di 15,72% e indica che il 15,72% della varianza dei dati può essere spiegata dalle variabili del modello.

Residui Peso

Min	Q1	Mediana	Q3	Max
-33.673	-9.574	-1.896	7.620	62.056

Analizzando la descrizione dei residui del modello, possiamo notare che il valore minimo e massimo dei residui sono abbastanza elevati, rispettivamente -33.673 e 62.056, il che suggerisce che ci potrebbero essere alcuni valori anomali o outliers presenti nei dati. Inoltre, il valore della mediana è vicino a zero indicando che la maggior parte dei residui si concentra attorno allo zero. La distribuzione dei residui sembra essere leggermente asimmetrica, poiché il valore della mediana non corrisponde esattamente al valore medio.

d)

Per confrontare il peso di Jane previsto dal modello con quello effettivo possiamo sostituire i valori di altezza, genere ed età nell'equazione di regressione e ottenere:

Peso previsto = -54,3 + 0,7633 * 165 - 3,99 * 1 + 0,2420 * 68 = 84,1 kg

Essendo il peso effettivo di 55 kg, Jane pesa meno di quanto previsto dal modello. Se calcoliamo l'indice di massa corporea di Jane (IMC):

$$\text{IMC} = \frac{peso}{altezza^2} = \frac{55}{1.65^2} = 20.20$$

Jane risulta in normopeso (18.5 - 24.9) e sarebbe, secondo il modello, in sovrappeso con un valore di IMC di 30,91 quindi il peso previsto differisce di molto da quello effettivo, pertanto, non è considerabile un modello adeguato. In generale i dati di altezza, genere ed età potrebbero non essere sufficienti per una valutazione dello stato di peso di una persona, in quanto bisogna tener conto anche della composizione corporea e dello stile di vita.

e)

Il modello finale scelto ha come equazione:

altezza= 144,57 + 12,228 genere + 0,1641 peso - 0,1298 età

Coefficienti

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	144,57	1,79	80,81	0,000	
genere	12,228	0,585	20,92	0,000	1,09
peso	0,1641	0,0183	8,96	0,000	1,04
eta_	-0,1298	0,0263	-4,93	0,000	1,06

Questo modello presenta il livello di R^2 aggiustato più alto con tutte le variabili esplicative che hanno un ottimo livello di significatività. Il modello di regressione lineare multipla utilizza le variabili indipendenti peso, genere ed età per prevedere il valore della variabile dipendente altezza. Il valore di 144,57 rappresenta l'altezza prevista quando tutte le altre variabili sono uguali a zero. Il coefficiente associato alla variabile altezza indica che un aumento di 1 kg di peso è associato ad un aumento di 0,16 cm all'altezza prevista (144,57), mantenendo costanti le altre variabili. Il coefficiente associato alla variabile genere indica che, mantenendo costanti le altre variabili, gli individui di genere 2 hanno un'altezza mediamente maggiore di 12 cm rispetto a quelli di genere 1. Il coefficiente associato alla variabile età indica che, mantenendo costanti le altre variabili, un aumento di 1 anno di età è associato ad una diminuzione di 0.13 cm.
Per ogni variabile indipendente il valore del P-value con un α=0,1 è significativo.
Dal modello si evidenzia che la variabile indipendente genere è quella che incide di più sulla variabile dipendente altezza. Tutte le variabili hanno valori VIF inferiori a 2, e questo indica che la correlazione tra le variabili indipendenti è relativamente bassa e che il modello è stabile.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
6,56652	55,36%	55,12%	54,66%

Il valore di R-quadrato aggiustato del modello è di 55,36% e indica che il 55,36% della varianza dei dati può essere spiegata dalle variabili del modello.

f)

Residui Altezza

Min	Q1	Mediana	Q3	Max
-17.59	-4.53	0.167	4.51	19.46

Analizzando la descrizione dei residui del modello possiamo notare che il valore minimo e massimo dei residui, rispettivamente -17.59 e 19.46, non sono abbastanza elevati, il che suggerisce che ci potrebbero essere pochi valori anomali o outliers presenti nei dati. Inoltre, il valore della mediana è vicino a zero indicando che la maggior parte dei residui si concentra attorno allo zero. La distribuzione dei residui sembra essere leggermente asimmetrica, poiché il valore della mediana non corrisponde esattamente al valore medio.

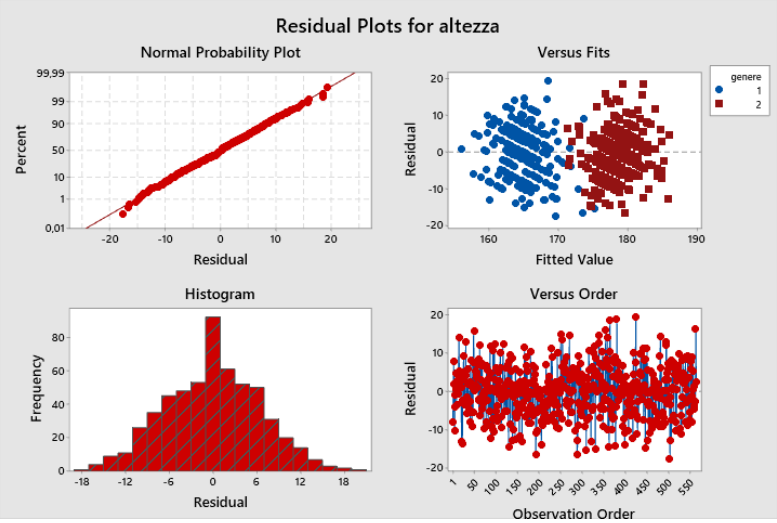


Figura 6 con Minitab

Nel primo grafico della Figura 6 in alto a sinistra è evidente che la distribuzione dei residui è normale e quasi tutti punti si trovano sulla retta. Nel secondo grafico in alto a destra la distribuzione dei residui risulta poco lineare con elevata presenza di molti valori sopra e sotto la linea dello zero e c'è la presenza di due gruppi di osservazioni per via della diversificazione dei generi. La terza figura in basso a sinistra è un istogramma avente nell'asse delle ordinate le frequenze dei residui con un picco nel valore 0 con una frequenza maggiore di 80.
Nel quarto grafico in basso a destra ha sull'asse delle ordinate i residui e sull'asse delle ascisse l'ordine in cui i dati sono stati raccolti, e si nota che le distribuzioni delle osservazioni sono casuali quindi il modello lineare si può considerare buono.

g)

L'indagine effettuata su peso, altezza ed età per un campione di 565 persone di provenienza geografica non definita di età tra i 16-72 anni risulta approssimativa. I dati non sono completi. Sarebbe stato utile conoscere area geografica di provenienza, occupazione, stile di vita (fumo, consumo di alcool, attività fisica) e abitudini alimentari degli individui esaminati.
Inoltre, l'andamento dei dati può essere influenzato dalla presenza di persone tra i 16-18 anni che hanno o non hanno raggiunto l'altezza definitiva, per contro con l'avanzare dell'età vi è una lenta riduzione della statura. Riguardo le donne vi sono fattori ormonali che influiscono sul peso.

Esercizio 2

a)

Il nuovo dataset è composto da 96 osservazioni per ogni dato di cui tre variabili categoriche: densità, fertilizzante usato per la piantagione e il blocco di substrato di terreno e da una variabile quantitativa resa del mais. La resa del mais maggiore è di 179,06 kg/m^2 proveniente dal substrato di terreno 4, densità di piantagione 2 (fitto) e somministrazione di fertilizzante 3, mentre la resa minore è pari a 175,36 kg/m^2 con blocco di trattamento 1, densità di piantagione 1(rado) e uso del fertilizzante 1.
La differenza tra il valore minimo e il valore massimo è lieve con un valore di 3,7 kg/m^2.

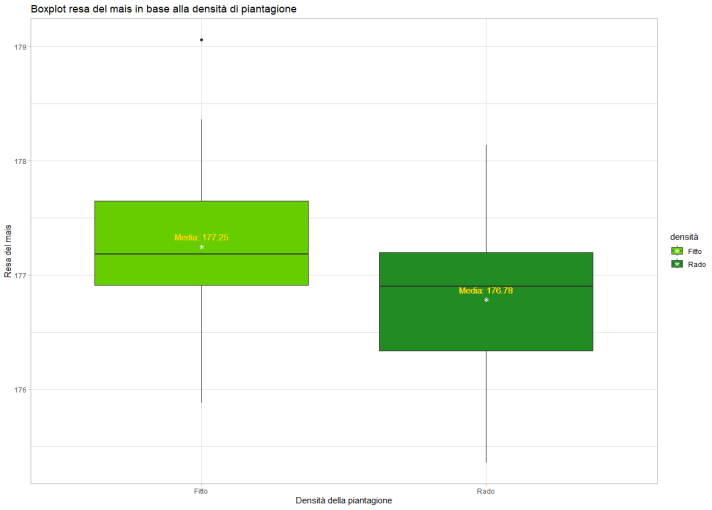


Figura 7

	Resa densità 1	Resa densità 2
Numero di osservazioni	48	48
Media	176,68	177,25
Deviazione standard	0,61	0,64
Minimo	175,36	175,88
Massimo	178,14	179,06
Primo quartile	176,32	176,90
Mediana (secondo quartile)	176,90	177,18
Terzo quartile	177,21	177,66
Simmetria/ Asimmetria	Asimmetria positiva in quanto (Q3 - Q2) > (Q2 - Q1)	Lieve asimmetria negativa in quando (Q3-Q2) < (Q2-Q1)

Tabella 2

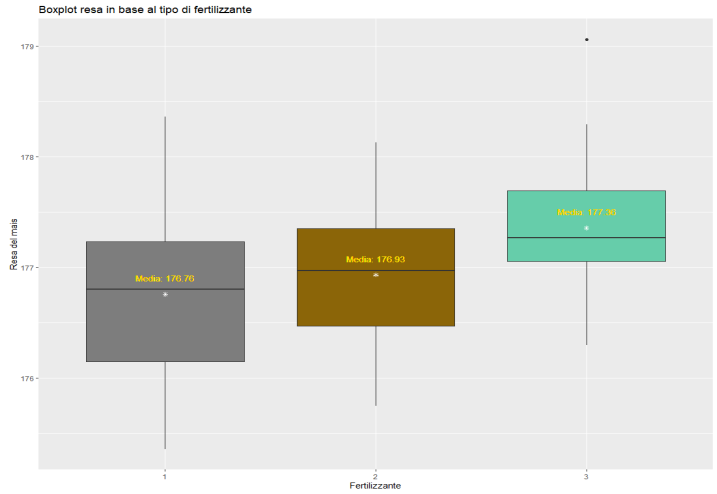


Figura 8

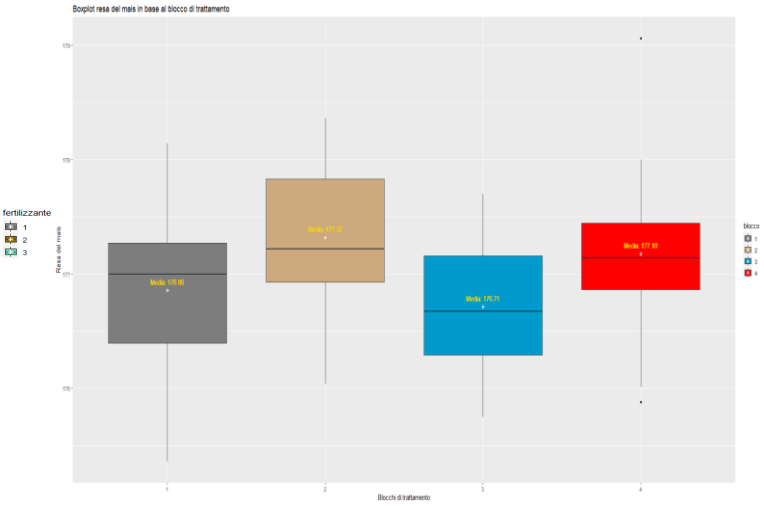


Figura 9

	Resa fertilizzante 1	Resa fertilizzante 2	Resa fertilizzante 3	Resa blocchi di trattamento 1	Resa blocchi di trattamento 2	Resa blocchi di trattamento 3	Resa blocchi di trattamento 4
Numero di osservazioni	32	32	32	24	24	24	24
Media	176,76	176,93	177,36	176,86	177,32	176,71	177,18
Deviazione standard	0,69	0,57	0,60	0,63	0,65	0,59	0,65
Minimo	175,36	175,75	176,30	175,36	176,04	175,75	175,88 (outlier)
Massimo	178,36	178,13	179,06 (outlier)	178,14	178,36	177,70	179,06 (outlier)
Primo quartile	176,14	176,46	177,70	176,36	176,93	176,23	176,85
Mediana (secondo quartile)	176,80	176,97	177,26	177	177,22	176,88	177,14
Terzo quartile	177,23	177,36	177,70	177,31	177,88	177,17	177,48
Simmetria/ Asimmetria	Lieve asimmetria negativa in quanto (Q3-Q2) <(Q2-Q1)	Quasi simmetrico	Asimmetria negativa in quanto (Q3-Q2) >(Q2-Q1)	Asimmetria negativa in quanto (Q3-Q2) < (Q2-Q1)	Asimmetria positiva in quanto (Q3-Q2) < (Q2-Q1)	Quasi simmetrica	Simmetrico

Tabella 3

b,c)

Per valutare l'impatto del fertilizzante sulla resa, si crea un modello di regressione lineare multipla avente come variabile dipendente la resa e come variabile indipendente il fertilizzante

Regression Equation

resa = 176,416 + 0,2995 fertilizzante

Coefficienti

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	176,416	0,168	1052,32	0,000	
fertilizzante	0,2995	0,0776	3,86	0.000208	1,00

Il coefficiente fisso di 176,416 rappresenta la resa prevista quando tutte le altre variabili sono uguali a zero. Il modello di regressione lineare multipla stima che per ogni aumento nel livello del fertilizzante il coefficiente fisso della resa aumenta di 0.2995 kg/m^2 e il valore del P-Value è molto basso (0.000208) quindi la variabile indipendente fertilizzante è altamente significativa per il modello.

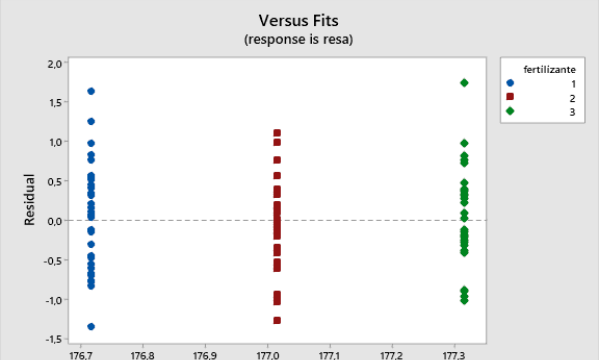


Figura 10 con Minitab

La Figura 10 ci mostra che i tre tipi di fertilizzante non presentano significative differenze. In particolare, il fertilizzante 3 presenta un valore outlier di 179,6 kg/m^2 che fa contrarre la media verso l'alto portandola a 177,36 kg/m^2 rispetto al fertilizzante 1 in media 176,76 e fertilizzante 2 in media 176,93.

d)

Per valutare l'impatto del fertilizzante e la densità del suolo sulla resa, si crea un modello di regressione lineare multipla avente come variabile dipendente la resa e come variabili indipendenti il fertilizzante e la densità.

Regression Equation

resa = 175,724+ 0,2995 fertilizzante+ 0,462 densità

Coefficienti

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	175,724	0,236	744,20	0,000	
fertilizzante	0,2995	0,0723	4,14	0,000	1,00
densità	0,462	0,118	3,91	0.000174	1,00

Il coefficiente fisso di 175,4724 ha la resa prevista quando tutte le altre variabili sono uguali a zero. Mantenendo costante la densità del suolo, un aumento di una unità nella quantità di fertilizzante utilizzato è associato a un aumento di 0.2995 kg /m^{2}del coefficiente fisso della resa. Allo stesso modo, mantenendo costante la quantità di fertilizzante utilizzato un aumento di una unità nella densità del suolo è associato a un aumento di 0.4619 kg/m^2 del coefficiente fisso della resa. Entrambe le variabili indipendenti mostrano un ottimo livello di significatività della variabile dipendente resa. Il valore del R^2 aggiustato è pari a 0,2428 ciò significa che il modello spiega il 24% della varianza nella resa mentre il restante 76% della varianza rimane ancora inspiegato dal modello.

e,f)

Per valutare la presenza di una possibile interazione tra il fertilizzante e la densità del suolo sulla resa, si crea un modello di regressione lineare multipla avente come variabile dipendente la resa e come variabili indipendente l'interazione tra fertilizzante e densità.

Regression Equation

resa = 175,434 - 0,097 densità*fertilizzante + 0,444 fertilizzante + 0,655 densità

Coefficienti

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	175,434	0,495	354,14	0,000	
densità*fertilizzante	-0,097	0,145	-0,67	0,507	16,00
fertilizzante	0,444	0,229	1,94	0,056	10,00
densità	0,655	0,313	2,09	0,039	7,00

Il fertilizzante e la densità presi singolarmente risultano ancora significativi per il modello ma la loro interazione non risulta significativa e pertanto concludiamo che non c'è interazione tra la densità e il fertilizzante sulla variabile dipendente resa. Il modello di regressione lineare multipla tra la variabile dipendente resa e le variabili indipendenti densità e fertilizzante risulta migliore.

g)

Per valutare l'impatto dei diversi substrati di terreno sulla resa si crea un modello di regressione lineare multipla tra le variabili indipendenti fertilizzante e la densità del suolo e i diversi substrati di terreno sulla variabile dipendente resa.

Regression Equation

resa = 175,795 + 0,533 densità - 0,0714 blocco + 0,2995 fertilizzante

Coefficienti

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	175,795	0,243	724,09	0,000	
densità	0,533	0,132	4,05	0,000	1,25
blocco	-0,0714	0,0589	-1,21	0,229	1,25
fertilizzante	0,2995	0,0721	4,15	0,000	1,00

Il coefficiente fisso di 175,795 rappresenta la resa prevista quando tutte le altre variabili indipendenti sono uguali a zero.

- 0,533 è il valore del coefficiente associato alla variabile indipendente densità mantenendo costanti le altre variabili: un aumento di un'unità di densità fa aumentare il coefficiente fisso della resa di 0,533 kg/m^2.
- -0,0714 è il valore del coefficiente associato alla variabile indipendente blocco mantenendo costanti le altre variabili: un aumento di un'unità di blocco di substrato di terreno fa diminuire il coefficiente fisso della resa di 0,07 kg/m^2.
- 0,2995 è il valore del coefficiente associato alla variabile indipendente fertilizzante mantenendo costanti tutte le altre variabili: un aumento di un'unità di fertilizzante usato fa aumentare il coefficiente fisso della resa di 0,299 kg/m^2.

Le variabili indipendenti densità e fertilizzante sono molto significative per il modello mentre la variabile indipendente blocco non lo è significativa con un α=0,1 pertanto si potrebbe escludere dal modello. Il modello di regressione multipla preso in analisi presenta un R^2 aggiustato pari a 0,2466, ciò significa che il modello spiega il 24,66% della varianza nella resa mentre il restante 75,34% della varianza rimane ancora inspiegato dal modello.

h)

Il modello di regressione lineare multipla scelto come modello finale ha la resa del mais come variabile dipendente e come variabili indipendenti il fertilizzante usato e la densità. È stato scelto questo modello perché il P-value di entrambe le variabili indipendenti risulta molto basso e quindi si ha un ottimo livello di significatività rispetto al modello con l'aggiunta della variabile indipendente blocco che non risultava significativo. Il valore del R^2 aggiustato è di 0,2428 e ciò significa che il modello spiega il 24% della varianza nella resa mentre il restante 76% della varianza rimane ancora inspiegato dal modello. Questo modello ha il valore del Cp di Mallows di 3,5 rispetto al modello di regressione lineare con la variabile blocco con un Cp di Mallows di 4.

Min	Q1	Mediana	Q3	Max
-1.20646	-0.36849	-0.03169	0.34292	1.51401

La distanza tra il valore minimo e il valore massimo dei residui della resa del mais non è alta e la mediana è molto vicina allo 0.

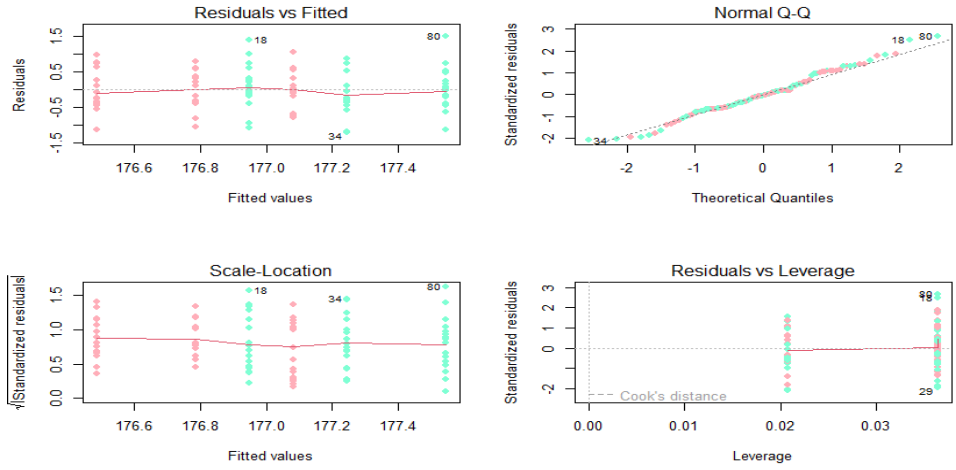


Figura 11

La Figura 11 mostra quattro tipi di grafici dei residui della resa del mais.
Il primo grafico in alto a sinistra mostra una dispersione dei punti intorno allo zero e alcuni punti sopra e sotto lo 0, in particolare il valore anomalo di 179,6 kg/m^2 dell'osservazione 80 e il valore di 178.36 kg/m^2 dell'osservazione 18 e il valore di 176.04 kg/m^2 dell'osservazione 34. Il secondo grafico in alto a destra mostra che la distribuzione dei residui ha una distribuzione normale con l'eccezione di alcuni valori, tra cui le osservazioni 80,18,34. Il terzo grafico in basso a destra mostra che la radice quadratica dei residui standardizzati dovrebbero variare tra 0 e $\sqrt{2}$.
Il quarto grafico in basso a destra mostra sull'asse delle ordinate i valori standardizzati dei residui e sull'asse delle ascisse il valore del leverage ovvero una misura fornisce l'informazione su quali sono i valori influenti. In questo grafico non c'è nessuna osservazione che superi la distanza di Cook di D_i>0,5 e quindi non c'è nessuna osservazione che sia troppo influente.