Corso di Statistica per Ingegneria Fisica
Anno accademico 2020/2021

Esercitazione 10 - Regressione lineare

Esercizio 1. In una certa comunità si registrano mensilmente il consumo di gelati X misurato in kg e il numero di casi di allergia al polline Y. I dati raccolti nell'ultimo anno forniscono le seguenti informazioni:

$$\bar{x} = 110 \qquad s_X = 100$$
$$\bar{y} = 20 \qquad s_Y = 5$$

Il coefficiente di correlazione vale 0.93.

(a) Si determini la retta di regressione di Y su X e se ne disegni il grafico. *[$\hat{y} = 0.0465\,x + 14.885$]*

(b) Si stimi (puntualmente) il numero di casi di allergia in un mese in cui il consumo di gelati è pari a 300 kg. La previsione ottenuta è buona?
[$\hat{y}^* = 0.0465 \cdot 300 + 14.885 = 28.835$. La previsione sembra buona, perché il modello lineare spiega bene la variabilità della risposta ($|r| = 0.93 > 0.9$); però dovrei guardare i residui per verificare che siano soddisfatte le ipotesi di gaussianità.]

(c) Mediante un intervallo di confidenza di livello 90% si stimi il numero atteso di casi di allergia in un mese in cui il consumo di gelati è pari a 300 kg. Si suppongano valide le ipotesi del modello lineare gaussiano.
[$28.835 \pm 1.8125 \cdot \sqrt{3.71525} \cdot \sqrt{\frac{1}{12} + \frac{(300-110)^2}{110000}} = 28.835 \pm 2.241$]

(d) Mediante un intervallo di previsione di livello 90% si stimi il numero di casi di allergia in un mese in cui il consumo di gelati è pari a 300 kg. Si suppongano sempre valide le ipotesi del modello lineare gaussiano.
[$28.835 \pm 1.8125 \cdot \sqrt{3.71525} \cdot \sqrt{1 + \frac{1}{12} + \frac{(300-110)^2}{110000}} = 28.835 \pm 4.151$]

Esercizio 2. La Takeo General Company sta finendo ora di produrre le 50 unità del sofisticatissimo componente TG7 che le furono commissionate dal Governo l'anno scorso. Poiché il Governo ha appena richiesto di sottoporgli una proposta di contratto per la produzione di altre 250 unità di TG7, la Takeo General Company desidera studiare la relazione fra le ore di lavoro necessarie per produrre un TG7 e il numero di TG7 già prodotti. Infatti all'aumentare delle unità prodotte X, aumenta l'esperienza nella produzione e di conseguenza diminuiscono le ore di lavoro richiesto. I dati vengono trasformati per eseguire una regressione lineare di $\log Y$ su $\log X$. Considerando l'output di R riportato sotto, si risponda alle seguenti domande.

(a) Si espliciti il legame supposto fra Y e X nell'elaborazione del modello di regressione.
[$\log Y = \beta_0 + \beta_1 \log x + E$ con $E \sim N(0, \sigma^2)$]

(b) Si scriva l'equazione di regressione stimata. *[$\log \hat{y} = 6.60992 - 0.55784 \log x$]*

(c) Si giudichi la bontà del modello.
[Il modello sembra buono a giudicare dall'R^2 elevato e dalla significatività del regressore $\log X$; bisognerebbe però esaminare anche i residui.]

(d) Si deduca da (b) l'equazione stimata che dà Y in funzione di X e se ne disegni un grafico qualitativo.
[$\hat{y} = \mathrm{e}^{6.60992}\,(1/x)^{0.55784}$]

Si vuole ora inferire sulle ore di lavoro necessarie per produrre il centesimo TG7 delle possibili ulteriori 250 unità (si è quindi interessati al caso $x = 50 + 100 = 150$).

(e) Si fornisca una previsione puntuale per Y. *[$\hat{y} = 45.367$]*

(f) Si fornisca una previsione intervallare per $\log Y$ al 99%.
[$3.8148 \pm 2.6822 \cdot 0.1503 \cdot \sqrt{1 + \frac{1}{50} + \frac{(\log 150 - 2.7302)^2}{2.8635}} = 3.8148 \pm 0.6789$]

(g) Si deduca da (f) una previsione intervallare per Y al 99%. *[$(23.009, 89.452)$]*

(h) Si fornisca una stima intervallare per il valore medio di $\log Y$ ad un livello di confidenza del 99%.
[$3.8148 \pm 2.6822 \cdot 0.1503 \cdot \sqrt{\frac{1}{50} + \frac{(\log 150 - 2.7302)^2}{2.8635}} = 3.8148 \pm 0.5463$]

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.60992    0.24343  27.153 0.000110 ***
log(x)      -0.55784    0.08882  -6.281 0.008150 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1503 on 48 degrees of freedom
Multiple R-squared: 0.9293,    Adjusted R-squared: 0.9058
F-statistic: 39.45 on 1 and 48 DF,  p-value: 0.00815
```

Esercizio 3. La $CMMR$ è interessata allo studio dell'effetto dei diversi livelli di acidità del suolo (X) sul prodotto (Y) di un certo raccolto. Uno degli obiettivi dello studio è lo sviluppo di un modello utilizzabile per prevedere il prodotto quando sono presenti diversi livelli di acidità del suolo. La tabella allegata mostra il diagramma di dispersione dei dati ottenuti da 20 diversi suoli (Fig. 1) e l'elaborazione di questi dati sulla base di due modelli di regressione lineare: Y su X (Modello 1, Fig. 2) e Y su X e X^2 (Modello 2, Fig. 3). In entrambi i casi è fornito anche il grafico dei residui standardizzati.

(a) Si espliciti il legame fra Y, X ed X^2 per i due modelli considerati, stimando i parametri che in essi compaiono, possibilmente mediante intervalli di confidenza al 95%.
[Modello 1: $Y = \beta_0 + \beta_1 x + E$, con $E \sim N(0, \sigma^2)$; $\beta_0 = -7.6311 \pm 2.1009 \cdot 4.0462 = -7.6311 \pm 8.5007$, $\beta_1 = 12.1039 \pm 2.1009 \cdot 0.6886 = 12.1039 \pm 1.4467$.
Modello 2: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + E$, con $E \sim N(0, \sigma^2)$; $\beta_0 = 5.1399 \pm 2.1098 \cdot 4.2431 = 5.1399 \pm 8.9521$, $\beta_1 = 4.2748 \pm 2.1098 \cdot 1.9462 = 4.2748 \pm 4.1061$, $\beta_2 = 0.7752 \pm 2.1098 \cdot 0.1863 = 0.7752 \pm 0.3931$.]

Si confrontino i due modelli.

(b) Qual è da considerarsi migliore? Perché?
[Entrambi i modelli hanno R^2_{adj} elevato; quello del Modello 2 ($r^2_{adj} = 0.9695$) è leggermente maggiore di quello del Modello 1 ($r^2_{adj} = 0.9419$). Entrambi i modelli sono globalmente significativi (p-value dell'F-test $= 8.827 \cdot 10^{-13}$ per il Modello 1, $= 5.046 \cdot 10^{-14}$ per il Modello 2). Tuttavia, l'ipotesi del modello lineare gaussiano non sembra essere soddisfatta per il Modello 1; i suoi residui non sono infatti omoschedastici, e mostrano invece un pattern a U chiaramente riconoscibile. Ciò è indice di una probabile dipendenza da X^2 non colta dal modello. I residui del Modello 2 sono invece disposti a nuvola e non sembrano mostrare pattern particolari; sarebbe utile però sottoporli a un ulteriore test di normalità. Nel Modello 2, inoltre, entrambi i predittori X e X^2 sono significativi (p-value del test sul primo predittore $= 0.042220$, del test sul secondo predittore $= 0.000654$). Alla luce del fatto che il Modello 2 sembra soddisfare le ipotesi del modello lineare gaussiano ed entrambi i suoi predittori sono significativi, lo preferiamo al Modello 1.]

Si è ora interessati al caso di acidità 4.

(c) Si preveda puntualmente il corrispondente prodotto del raccolto.
[$\hat{y} = 5.1399 + 1.9462 \cdot 4 + 0.7752 \cdot 4^2 = 25.3279$.]

Modello 1:

```
Call: lm(formula = y ~ x)
Residuals:
     Min       1Q   Median       3Q      Max
-17.4758  -6.1522   0.8324   5.0530  21.1802
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.6311     4.0462  -1.886   0.0755 .
x            12.1039     0.6886  17.578  8.83e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.252 on 18 degrees of freedom Multiple
R-squared: 0.945,      Adjusted R-squared: 0.9419
F-statistic:   309 on 1 and 18 DF,  p-value: 8.827e-13
```

Modello 2:

```
Call: lm(formula = y ~ x + I(x^2))
Residuals:
      Min        1Q    Median        3Q       Max
-10.95173  -5.34724  -0.08435   2.42665  13.09173

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1399     4.2431   1.211 0.242330
x             4.2748     1.9462   2.196 0.042220 *
I(x^2)        0.7752     0.1863   4.162 0.000654 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.701 on 17 degrees of freedom Multiple
R-squared: 0.9727,     Adjusted R-squared: 0.9695
F-statistic: 303.2 on 2 and 17 DF,  p-value: 5.046e-14
```

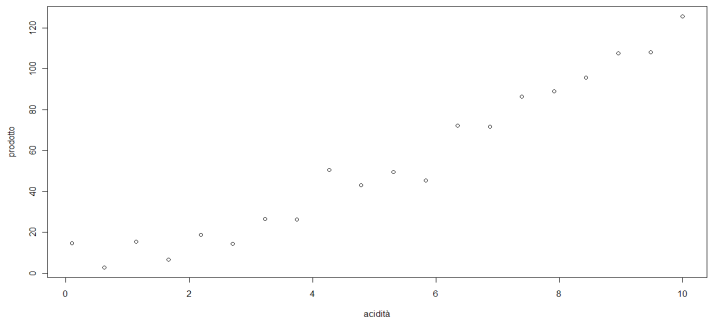


Figura 1: Esercizio 3: scatterplot dei dati

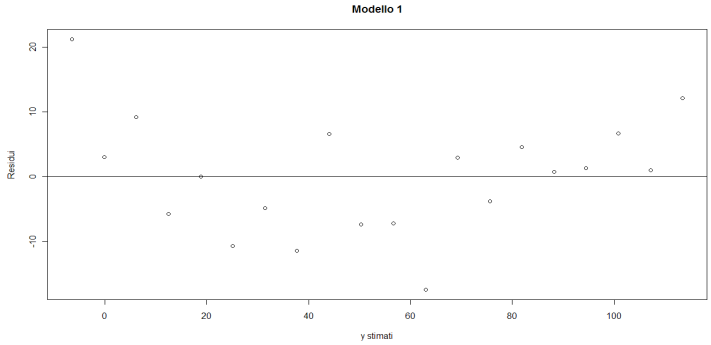


Figura 2: Esercizio 3: residui del Modello 1

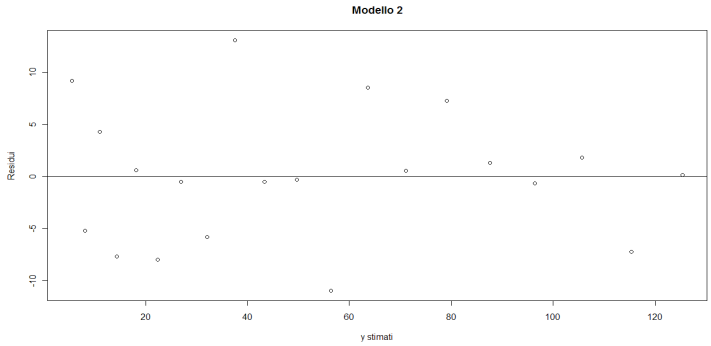


Figura 3: Esercizio 3: residui del Modello 2

Esercizio 4. Vengono effettuate 20 misurazioni della lunghezza di una molla sottoposta a vari pesi. La lunghezza y è misurata in pollici ed il peso x in libbre. Secondo la legge di Hooke, la lunghezza della

molla è data dalla relazione

$$y = \beta_0 + \beta_1 x$$

dove β_0 è la lunghezza della molla quando questa non è sottoposta a pesi e β_1 è la sua costante di elasticità. Si assume dunque il modello

$$Y = \beta_0 + \beta_1 x + E$$

dove E è l'errore di misurazione (distribuito normalmente con media nulla). Si svolgono infine le analisi dei dati.

(a) Completare l'output R dell'analisi riportato in Fig. 4. *[Vedi Fig. 13 in fondo al file.]*

(b) Discutere la bontà del modello utilizzando le informazioni riportate nell'output in Fig. 4 e 5. *[Il modello lineare spiega molto bene la variabilità dei dati ($r^2 = 90.18\%$). Le ipotesi del modello lineare gaussiano sembrano essere rispettate, come si vede dall'omoschedasticità dei residui. Sarebbe però opportuno avere a disposizione anche lo scatterplot dei residui standardizzati per individuare eventuali outlier, e il loro normal qq-plot e/o il p-value del test di Shapiro-Wilks per verificare meglio la loro gaussianità. Infine, l'unico predittore è estremamente significativo ($p\text{-value} < 1.65 \cdot 10^{-10}$).]*

(c) Fornire un intervallo di confidenza al livello del 95% per i parametri β_0 e β_1.
[$\beta_0 \in (4.96991 \pm 2.1009 \cdot 0.03805) = (4.96991 \pm 0.07994)$
$\beta_1 \in (0.20805 \pm 2.1009 \cdot 0.01618) = (0.20805 \pm 0.03399)$ *.]*

(d) Scrivere la retta di regressione stimata. Calcolare quanto vale l'allungamento medio della molla, stimato dal modello, per un aumento di 1 libbra del peso. *[*$\hat{y} = 4.96991 + 0.20805\, x$. $\widehat{\mathbb{E}[\Delta Y]} = 0.20805$*]*

(e) Fornire una stima puntuale ed un intervallo di previsione al 95% per la lunghezza della molla sottoposta ad un peso di 1.3 libbre.
[$\hat{y}^* = 4.96991 + 0.20805 \cdot 1.3 = 5.24038$.
$IP_{Y^*}(0.95) = \left(5.24038 \pm 2.1009 \cdot 0.064 \sqrt{1 + \frac{1}{20} + \frac{(1.3 - 2.17900)^2}{15.64599}}\right) = (5.24038 \pm 0.14098)$ *.]*

```
Call:
lm(formula = lunghezza ~ peso)

Residuals:
      Min        1Q    Median        3Q       Max
-0.097421 -0.055969  0.004998  0.053402  0.108915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.96991    0.03805  130.61  < 2e-16 ***
peso         0.20805    [illegible]   12.86 1.65e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.064 on 18 degrees of freedom
Multiple R-squared: 0.9018,     Adjusted R-squared: 0.8963
F-statistic: [illegible] on [illegible] and [illegible] DF,  p-value: [illegible]
```

Figura 4: Esercizio 4: output analisi con informazioni mancanti

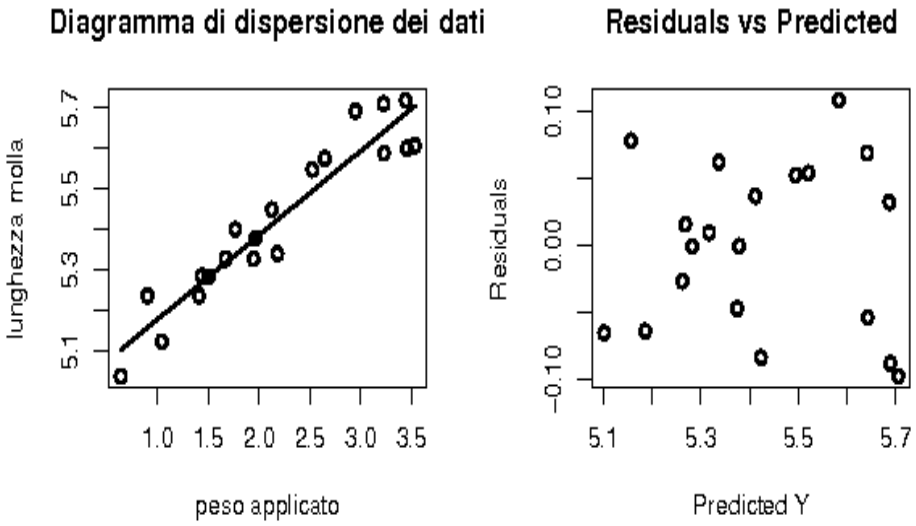


Figura 5: Esercizio 4: diagramma di dispersione dei dati e grafico dei residui

Esercizio 5. Una società di ricerche di mercato vuole prevedere la tiratura di quotidiani durante il fine settimana (Y) in diverse aree di mercato, in migliaia di copie. La società seleziona come variabile indipendente la densità di popolazione (X), in abitanti per Km^2. La tabella allegata riassume i risultati ottenuti da una serie di 25 misure di X e dei valori di Y corrispondenti, e la loro elaborazione sulla base di un modello di regressione lineare.

(a) Si completi l'output dell'analisi mostrato in Fig. 6. *[Vedi Fig. 14 in fondo al file.]*

(b) Qual è l'equazione di regressione stimata? *[$\hat{y} = 0.550032 + 0.023149\,x$]*

(c) Il modello lineare spiega bene la variabilità dei dati? *[$r^2 = 0.9434 > 0.81 \Rightarrow$ Sì.]*

Supponiamo d'ora in poi che siano verificate le ipotesi del modello lineare gaussiano.

(d) È significativo utilizzare X come variabile esplicativa per prevedere Y? Si risponda tramite un opportuno test ai livelli di significatività 1% e 5%.
[Sì, sia all'1% che al 5% (il test per le ipotesi $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ ha p-value $= 7.63 \cdot 10^{-16}$).]

(e) Ovviamente, il modello dovrebbe prevedere una tiratura media pari a 0 nelle aree con una densità di popolazione $x^* = 0$. I dati sono compatibili con tale requisito?
[No (il test per le ipotesi $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 \neq 0$ ha p-value $= 0.000153 \Rightarrow$ si rifiuta H_0 a tutti i livelli di significatività ragionevoli).]

(f) Fino all'anno scorso, l'andamento di Y in funzione di X era ben approssimato da una retta con pendenza pari a 0.025. C'è evidenza che quest'anno tale pendenza sia diminuita?
[No (il test per le ipotesi $H_0 : \beta_0 = 0.025$ vs. $H_1 : \beta_0 < 0.025$ ha $5\% <$ p-value $< 10\% \Rightarrow$ si accetta H_0 a tutti i livelli di significatività tipici).]

Ora si vuole utilizzare il modello elaborato per fare previsioni e stime sulla tiratura di quotidiani durante il fine settimana nelle aree dove la densità di popolazione è pari a $x^* = 65$ persone al chilometro quadrato.

(g) Si fornisca una stima puntuale della tiratura attesa in una qualunque delle aree considerate.
[$\hat{y}^* = 0.550032 + 0.023149 \cdot 65 = 2.054717$]

(h) Si fornisca un intervallo di confidenza al 95% per la tiratura attesa in una qualunque delle aree considerate.
[$2.054717 \pm 2.0687 \cdot 0.1629 \cdot \sqrt{\frac{1}{25} + \frac{(65-99.112)^2}{18993.6}} = 2.054717 \pm 0.107237$]

(i) Si fornisca un intervallo di previsione al 95% per la tiratura in una qualunque delle aree considerate.
[$2.054717 \pm 2.0687 \cdot 0.1629 \cdot \sqrt{1 + \frac{1}{25} + \frac{(65-99.112)^2}{18993.6}} = 2.054717 \pm 0.353642$]

```
Call:
lm(formula = Y ~ X)

Residuals:
      Min        1Q    Median        3Q       Max
-0.520594 -0.081497  0.006366  0.124812  0.246961

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.550032   0.121596   4.523 0.000153 ***
X           [illegible]   0.001182  19.586 7.63e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1629 on 23 degrees of freedom
Multiple R-squared: [illegible]     Adjusted R-squared: [illegible]
F-statistic: 383.6 on [illegible] and [illegible] DF,  p-value: 7.629e-16
```

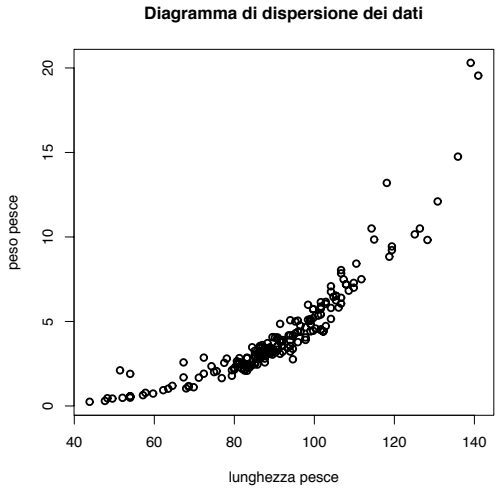
Figura 6: Esercizio 5: output analisi

Esercizio 6. Il pesce spatola (Polydon spathula) è un pesce d'acqua dolce comune nel Nord America. Documentazioni fossili testimoniano che esiste da almeno 300 milioni di anni. Anche se protetto da leggi in alcuni stati, il pesce spatola sta diventando una specie a rischio, vittima di eccessiva pesca, inquinamento e bracconaggio. Questi dati, raccolti nel 1970 nel fiume Mississippi, rappresentano una delle più grandi indagini statistiche sulla specie del pesce spatola allo scopo di studiare la dipendenza tra peso e lunghezza del pesce.

Si vuole dunque costruire un modello empirico per porre in relazione queste due grandezze: lunghezza, cioè la lunghezza del pesce (cm), e peso, cioè il peso del pesce (kg); per comodità chiamiamo le due variabili rispettivamente X e Y. Si propongono due modelli di regressione lineare, Y su X (Modello 1) e Y su X e X^2 (Modello 2), ipotizzano in prima istanza valide le ipotesi gaussiane.

(a) Esplicitare i due modelli empirici.
[Modello 1: $Y = \beta_0 + \beta_1 x + E,\ E \sim N(0, \sigma^2)$; Modello 2: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + E,\ E \sim N(0, \sigma^2)$.]

I dati raccolti costituiscono il campione di 185 osservazioni (x_i, y_i) rappresentato dal seguante diagramma di dispersione.



e i dati vengono elaborati secondo i due modelli proposti (vedi Fig. 7, 8, 9, 10).

(b) Si confrontino i due modelli, indicando quale si ritiene migliore.
[Entrambi i modelli risultano significativi, sia in termini di test F che di test sui singoli coefficienti; tuttavia, il Modello 2 presenta un R^2_{adj} più elevato ($\Rightarrow$ spiega meglio la variabilità delle Y), e mostra un grafico dei residui senza particolari anomalie ($\Rightarrow$ le ipotesi del modello lineare gaussiano sembrano rispettate, ma sarebbe utile un test di gaussianità sui residui). Il Modello 1, invece, ha R^2_{adj} più basso, e presenta un andamento nei residui che chiaramente indica la mancanza di un termine quadratico nel modello. Scegliamo dunque il Modello 2.]

(c) Per il modello scelto al punto (b) si esplicitino ipotesi nulla e alternativa del test di significatività dell'intercetta, il p-value dei dati osservati e si tragga una conclusione circa l'opportunità di eliminare l'intercetta dal modello. Che considerazione è possibile fare guardando il grafico dei residui?
[Test di significatività per l'intercetta: $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$; p-value$< 2*10^{-16}$, quindi rifiuto H_0: non elimino l'intercetta dal modello. I residui del modello presentano un possibile andamento eteroschedastico (variabilità che aumenta all'aumentare del valore predetto per Y)]

(d) Si stimi puntualmente il valore atteso di Y per $x^* = 27.25$. La stima è attendibile?
[$\hat{y}^* = 8.4666324 - 0.2658591 \cdot 27.25 + 0.0023371 * (27.25)^2 = 2.957415$. No, perché il valore $x^* = 27.25$ è troppo lontano dal range $[40, 140]$ delle x_i osservate.]

```
Call:
lm(formula = Peso ~ Lunghezza)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0967 -0.7540 -0.3896  0.2958  8.6981

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.501779   0.550197  -17.27   <2e-16 ***
Lunghezza    0.151749   0.005977   25.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.391 on 183 degrees of freedom
Multiple R-squared: 0.7789,     Adjusted R-squared: 0.7777
F-statistic: 644.6 on 1 and 183 DF,  p-value: < 2.2e-16
```

Figura 7: Esercizio 6: output modello 1

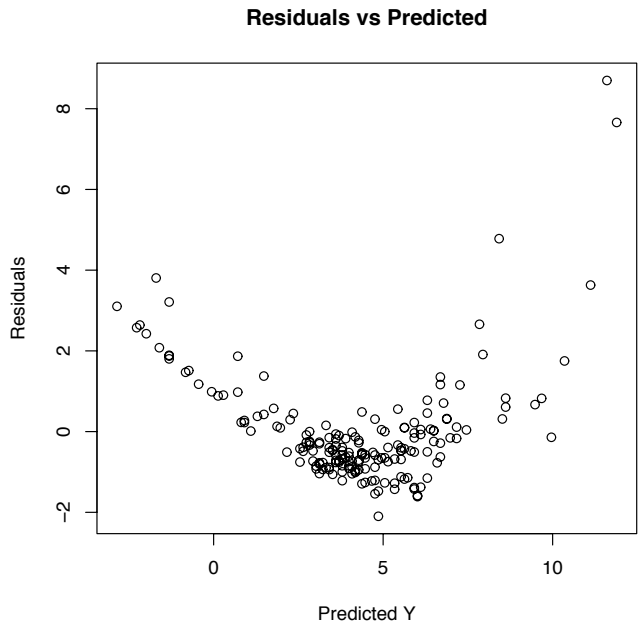


Figura 8: Esercizio 6: grafico dei residui modello 1

```
> LunghezzaSq <- Lunghezza^2
> result <- lm(Peso ~ Lunghezza + LunghezzaSq)
> summary(result)

Call:
lm(formula = Peso ~ Lunghezza + LunghezzaSq)

Residuals:
     Min       1Q   Median       3Q      Max
-2.99684 -0.32571 -0.04567  0.29645  3.60674

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.4666324  0.8999083   9.408   <2e-16 ***
Lunghezza   -0.2658591  0.0200069 -13.288   <2e-16 ***
LunghezzaSq  0.0023371  0.0001105  21.149   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7502 on 182 degrees of freedom
Multiple R-squared: 0.9361,     Adjusted R-squared: 0.9353
F-statistic:  1332 on 2 and 182 DF,  p-value: < 2.2e-16
```

Figura 9: Esercizio 6: output Modello 2

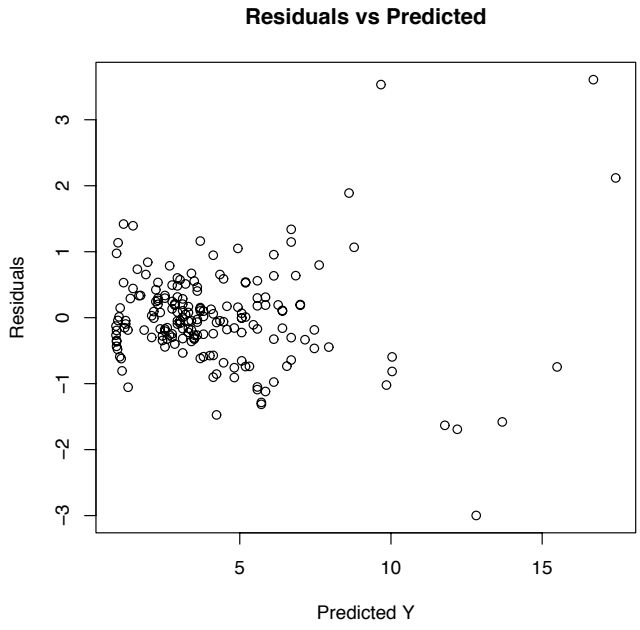
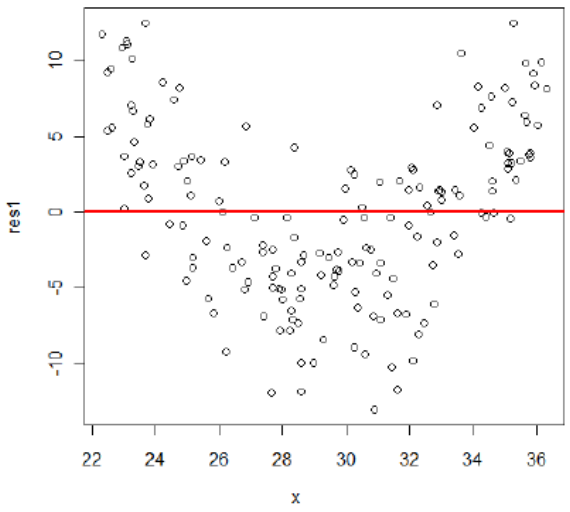


Figura 10: Esercizio 6: grafico dei residui Modello 2

Esercizio 7. Un tetro giorno d'autunno, due hobbit della Contea, Frodo e Sam, stanno girovagando allegramente nella foresta di Fangorn, quando si imbattono in una minacciosa impronta di orco. A tale vista, si interrogano subito su quanto possa pesare quell'essere spaventoso. Per rispondere a questa domanda, Frodo consulta il suo vecchio libro di statistica in cui si trovano i dati relativi a 172 orchi: per ciascuno di essi si ha il peso corporeo P (in kg) e la lunghezza dei piedi L (in cm). Frodo decide di spiegare la relazione fra queste due quantità impostando un modello empirico lineare gaussiano con responso P e predittore L.

(a) Scrivere la relazione ipotizzata fra P e L nel modello di Frodo.
[Modello Frodo: $P = \beta_0 + \beta_1 l + E \quad con \quad E \sim N(0, \sigma^2)$ *.]*

Per verificare la validità del suo modello, Frodo utilizza il software statistico R sul suo portatile, di cui riportiamo in Fig. 11 una sintesi dell'analisi e alcuni grafici dei residui del modello.




```
> summary(reg1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-13.070  -4.225  -0.051   3.697  12.526

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) [redacted]    3.2619 [redacted] [redacted]
x           [redacted] [redacted] [redacted]  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.804 on 170 degrees of freedom
Multiple R-squared: [redacted]      Adjusted R-squared: 0.8692
F-statistic:  1137 on 1 and 170 DF,  p-value: < 2.2e-16

> shapiro.test(res1)

        Shapiro-Wilk normality test

data:  res1
W = 0.9884, p-value = 0.1699
```

Figura 11: Esercizio 7: output del modello di Frodo

(b) Sapendo anche che per i 172 orchi catalogati

$$\sum_{i=1}^{172} l_i = 5074.109 \qquad \sum_{i=1}^{172} p_i = 18449.46 \qquad \sum_{i=1}^{172} (l_i - \bar{l})(p_i - \bar{p}) = 10368.18$$

$$\sum_{i=1}^{172} (l_i - \bar{l})^2 = 2806.916 \qquad \sum_{i=1}^{172} (p_i - \bar{p})^2 = 44024.5$$

si completi l'output di R in Fig. 11, riportando i calcoli effettuati.

[$\hat{\beta}_1 = \frac{\sum_i (l_i - \bar{l})(p_i - \bar{p})}{\sum_j (l_j - \bar{l})^2} = 3.6938\,, \qquad \hat{\beta}_0 = \bar{p} - \hat{\beta}_1 \bar{l} = -1.7051\,, \qquad \mathrm{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_i (l_i - \bar{l})^2}} = 0.1095\,,$

$t_0 = \frac{\hat{\beta}_0}{\mathrm{se}(\hat{\beta}_0)} = -0.523\,, \qquad t_1 = \frac{\hat{\beta}_1}{\mathrm{se}(\hat{\beta}_1)} = 33.718\,, \qquad p\text{-}value \simeq 2[1 - \Phi(|t_0|)] = 0.602\,,$

$r^2 = 1 - \frac{ss_e}{ss_t} = 1 - \frac{(172-2)\,\hat{\sigma}^2}{\sum_i (p_i - \bar{p})^2} = 0.8699$ *.]*

Sam decide invece di impostare un'altro modello empirico gaussiano, sempre con responso P, ma con due predittori: L e L^2.

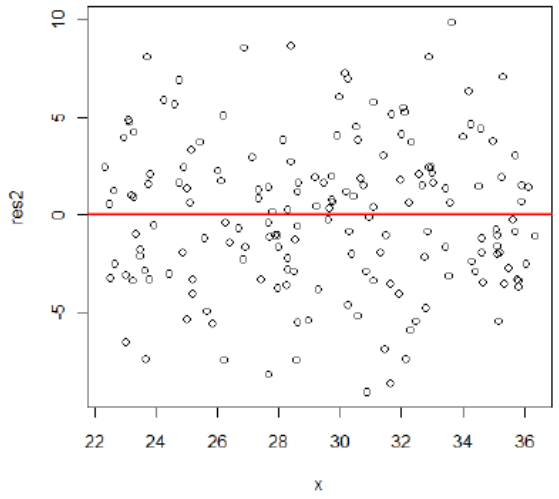
(c) Scrivere la relazione ipotizzata fra P e L nel modello di Sam.
[Modello Sam: $P = \beta_0 + \beta_1 l + \beta_2 l^2 + E \quad$ *con* $\quad E \sim N(0, \sigma^2)$ *.]*

In Fig. 12 riportiamo una sintesi dell'analisi e alcuni grafici dei residui del modello assunto da Sam. Naturalmente entrambi gli hobbit sostengono di aver creato il modello migliore.

(d) Aiutate Frodo e Sam a stabilire chi ha realizzato il modello migliore, giustificandone le ragioni.
[Il primo modello (realizzato da Frodo) presenta dei residui che non sembrano indipendenti dal predittore lunghezza dei piedi L: infatti essi tendono ad essere positivi per valori estremi di L e negativi per valori centrali di L. Le ipotesi gaussiane (= errori $E_1, \ldots, E_{172}$ i.i.d. normali a media nulla) non sembra verificata (si veda anche il relativamente basso p-value di Shapiro-Wilks: 0.1699*) e anzi il grafico dei residui suggerisce l'introduzione di L^2 fra i regressori. Il secondo modello invece (realizzato da Sam) presenta un grafico dei residui molto buono, con il classico andamento omoschedastico a nuvola, tipico nel caso di errori i.i.d., e un p-value di Shapiro-Wilks molto più alto:* 0.6596*. Inoltre in quest'ultimo modello i coefficienti sono tutti significativi e l'R^2 corretto è più alto.]*

Finalmente i due hobbit misurano l'impronta di orco trovata nella foresta: 33 cm.

(e) Aiutate Frodo e Sam a stimare (stima puntuale) il peso medio degli orchi con piedi lunghi quanto quello dell'impronta. *[Utilizzando il modello migliore:* $\hat{p} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 33 + \hat{\beta}_2 \cdot 33^2 = 119.3557$ *.]*



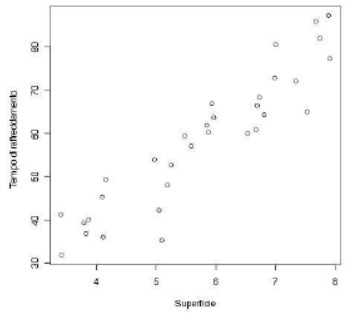

```
> summary(reg2)

Call:
lm(formula = y ~ x2 + x)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0704 -2.8262 -0.1425  2.1960  9.9133

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 239.61106   16.50892   14.51   <2e-16 ***
x2            0.28584    0.01939   14.74   <2e-16 ***
x           -13.07677    1.13973  -11.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.85 on 169 degrees of freedom
Multiple R-squared: 0.9431,     Adjusted R-squared: 0.9424
F-statistic:  1401 on 2 and 169 DF,  p-value: < 2.2e-16

> shapiro.test(res2)

        Shapiro-Wilk normality test

data:  res2
W = 0.9936, p-value = 0.6595
```

Figura 12: Esercizio 7: output del modello di Sam

Esercizio 8. Per i cilindri metallici prodotti dalla ABC, aventi tutti la stessa lunghezza, ma differente raggio, vogliamo studiare il tempo di raffreddamento Y (inteso come i minuti impiegati a tornare alla temperatura di 20°C dopo essere stati riscaldati a 50°C) in relazione alla superficie laterale X (variabile tra 3 e 8 cm^2). Assumiamo valido il modello gaussiano empirico

$$Y = \beta_0 + \beta_1 x + E \qquad E \sim N(0, \sigma^2)\,.$$

I dati relativi ad uno stock di $n = 32$ cilindri sono stati elaborati con R, ottenendo il diagramma di dispersione e l'output della regressione lineare qua riportati:




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6920     4.5288  -0.153     0.88
superficie   10.2198     0.7635  13.385 3.47e-14

Residual standard error:  6.031 on 30 degrees of freedom
Multiple R-squared:  0.8566, Adjusted R-squared:  0.8518
F-statistic:  179.2 on 1 and 30 DF, p-value:  3.474e-14
```

(a) Fornire una stima puntuale per la varianza degli errori E.
[Una stima puntuale del parametro σ^2 *è* $\hat{\sigma}^2 = (6.031)^2 = 36.373$ *.]*

(b) Fornire una previsione intervallare al 90% per il tempo di raffreddamento di una 33-esima sbarretta avente superficie laterale pari a $9\,\mathrm{cm}^2$.
$[IP_{Y(9)}(90\%) = \left(-0.6920 \pm 1.6973 \cdot 6.031 \cdot \sqrt{1 + \frac{1}{32} + \frac{(9-5.764924)^2}{62.396550}}\right)\, [80.0775\,,\, 102.4949]$ *.]*

(c) Verificare se possiamo ritenere $\beta_0 = 0$ mediante un opportuno test d'ipotesi, specificando: ipotesi nulla, ipotesi alternativa, regione critica di livello α, p-value dei dati raccolti, conclusione.
[Test per le ipotesi $H_0 : \beta_0 = 0$ *vs.* $H_1 : \beta_0 \neq 0$. $\mathrm{RC}(\alpha) = \{y_1, \ldots, y_n : |\hat{\beta}_0| > \mathrm{se}(\hat{\beta}_0)\, t_{1-\frac{\alpha}{2}}(30)\}$. p*-value* $= 0.88$. *Conclusione* debole*:* $\beta_0 = 0$ *agli usuali livelli di significatività.]*

(d) Verificare se possiamo ritenere $\beta_1 = 8$ mediante un opportuno test d'ipotesi, specificando: ipotesi nulla, ipotesi alternativa, regione critica di livello α, p-value dei dati raccolti, conclusione.
[Test per le ipotesi $H_0 : \beta_1 = 8$ *vs.* $H_1 : \beta_1 \neq 8$. $\mathrm{RC}(\alpha) = \{y_1, \ldots, y_n : |\hat{\beta}_1 - 8| > \mathrm{se}(\hat{\beta}_1)\, t_{1-\frac{\alpha}{2}}(30)\}$. $0.005 <$ p*-value* < 0.01 *(calcolato col computer:* p*-value* $= 0.0067939318$*). Conclusione* forte*:* $\beta_1 \neq 8$ *agli usuali livelli di significatività.]*

```
Call:
lm(formula = lunghezza ~ peso)

Residuals:
      Min        1Q    Median        3Q       Max
-0.097421 -0.055969  0.004998  0.053402  0.108915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.96991    0.03805  130.61  < 2e-16 ***
peso         0.20805    0.01618   12.86 1.65e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.064 on 18 degrees of freedom
Multiple R-squared: 0.9018,     Adjusted R-squared: 0.8963
F-statistic: 165.2 on 1 and 18 DF,  p-value: 1.654e-10
```

Figura 13: Esercizio 4: output analisi completo

```
Call:
lm(formula = Y ~ X)

Residuals:
      Min        1Q    Median        3Q       Max
-0.520594 -0.081497  0.006366  0.124812  0.246961

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.550032   0.121596   4.523 0.000153 ***
X           0.023149   0.001182  19.586 7.63e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1629 on 23 degrees of freedom
Multiple R-squared: 0.9434,     Adjusted R-squared: 0.941
F-statistic: 383.6 on 1 and 23 DF,  p-value: 7.629e-16
```

Figura 14: Esercizio 5: output analisi completo