6.4 Regressione lineare semplice

La *regressione lineare* è la parte della statistica che studia la relazione fra due o più variabili, che sono legate in modo NON DETERMINISTICO, per fare inferenze sul modello.

In particolare, si usano relazioni fra due o più variabili in modo da potere avere informazioni su una di esse conoscendo i valori dell'altra. Esempi di variabili che non sono legate fra loro da una relazione deterministica: x= l'età di un bambino e Y=la sua altezza, x=il volume di un motore e Y=il suo consumo di carburante, x=tempo di studio e Y= voto all'esame, ecc. Poiché x non è una variabile casuale la indichiamo con la lettera minuscola, mentre Y, che è un variabile casuale, viene indicata con la lettera maiuscola. Nel caso lineare supponiamo una relazione appunto lineare fra le due variabili x e Y:

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

. Questa relazione, di per sè deterministica, viene generalizzata a una relazione probabilistica. Date quindi informazioni su x e Y, l'obiettivo è quello di *predire* un valore futuro di Y per un particolare valore di x.

In questo modello, x viene detta *variabile indipendente* e Y viene detta *variabile dipendente*.

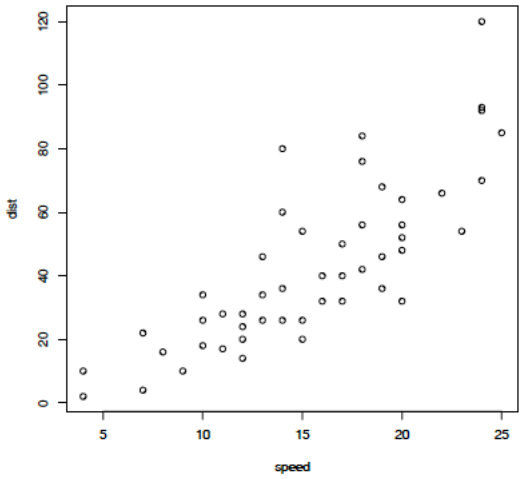
Il modello viene costruito a partire da alcune osservazioni $(x_i, Y_i), i = 1, \ldots n$..

Il grafico delle osservazioni individuate dalle coppie $(x_i, Y_i), i = 1, \ldots n$ viene detto *scatter plot*. Serve per dare un'idea della relazione esistente fra variabili x e Y.

Esempio 6.21 *Esempio di scatterplot con i dati del file* `cars` *del pacchetto* `datasets`*, che rappresentano distanze di arresto rispetto alle velocità di automobili (dati rilevati nel 1920).*

Soluzione.

```
> plot(dist~speed, data = cars)
```



6.4.1 Il modello lineare

Il modello deterministico più semplice fra due variabili x e Y è il modello lineare:

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

che graficamente rappresenta una retta il cui coefficiente angolare è β_1 e l'intercetta è β_0.

L'estensione al modello probabilistico è necessaria nel momento in cui le due variabili non hanno una relazione deterministica. In pratica, in corrispondenza di n variabili indipendenti $x_1, x_2, \ldots x_n$ si hanno n valori $Y_1, Y_2, \ldots Y_n$, che sono legati dalla relazione:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \;\; i = 1, \ldots n$$

quindi differiscono, rispetto al modello lineare esatto, di una quantità ϵ_i.

I valori Y_i sono in generale variabili aleatorie.

Formalmente, possiamo dare questa definizione del *modello di regressione lineare semplice*.

Modello di regressione lineare semplice. Esistono parametri $\beta_0, \beta_1, \sigma^2$ tali che, per ogni valore fissato della variabile indipendente x, la variabile dipendente è una variabile aleatoria legata ad x dal modello:

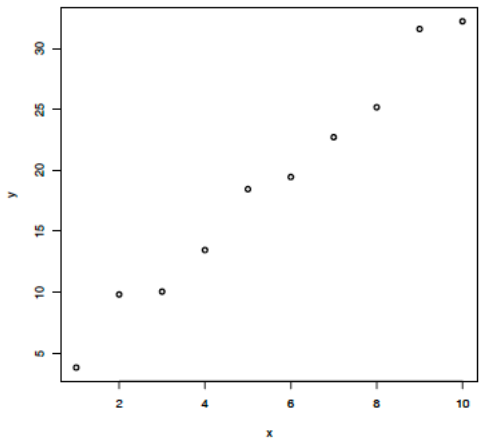
$$Y = \beta_0 + \beta_1 x + \epsilon,$$

dove ϵ è una variabile aleatoria, detta *errore casuale*, che si assume con distribuzione `norm(0,`σ`)`.

Esempio 6.22 *Esempio di scatterplot di dati.*

Soluzione.

```
> x <- 1:10
> v <- rnorm(10, 0, 1.5)
> y <- 2 + 3 * x + v
> plot(x, y)
```



Nella figura, si vede chiaramente che i punti non stanno *esattamente* sulla retta, ma c'è un piccolo errore ϵ_i per ogni punto.

6.4.2 Stima dei parametri

In questo paragrafo rispondiamo alla domanda:

'*Come calcolare stime dei parametri* β_0 *e* β_1 *della retta di regressione lineare assegnate le coppie* $(x_i, Y_i), i = 1, \ldots n$*? Cioé come determinare, fra le infinite rette del piano, una* buona *retta? Esiste una retta* migliore *delle altre?*'

Per rispondere, utilizziamo i concetti visti per le stime puntuali, visto che in pratica dobbiamo stimare i due parametri β_0 e β_1. In particolare, usiamo lo stimatore di massima verosimiglianza per calcolare le stime.

Visto che gli errori ϵ_i hanno distribuzione `norm(mean=0, sd=`σ`)`, allora la variabile aleatoria $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ ha distribuzione normale con deviazione standard σ. La funzione di verosimiglianza è:

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{n} f_{Y_i}(x_i) \\ &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} exp\left\{\frac{-(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-n/2} exp\left\{\frac{-\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\}; \end{aligned}$$

facendo il logaritmo naturale di $L(\beta_0, \beta_1)$ si ha:

$$F(\beta_0, \beta_1) = ln(L(\beta_0, \beta_1)) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}.$$

Per minimizzare questa funzione rispetto alle variabili β_0 e β_1:

$$\frac{\partial F}{\partial \beta_0} = 0, \quad \frac{\partial F}{\partial \beta_1} = 0.$$

Quindi:

$$\frac{\partial F}{\partial \beta_0} = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 x_i)(-1),$$

da cui:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i.$$

Per quanto riguarda l'altra derivata:

$$\begin{aligned} \frac{\partial ln L}{\partial \beta_1} &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 x_i)(-x_i) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i Y_i - \beta_0 x_i - \beta_1 x_i^2), \end{aligned}$$

da cui:

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i.$$

Quindi devo risolvere il sistema costituito dalle due seguenti equazioni:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i$$

che dà come soluzione:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}$$

e

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

dove $\bar{Y}$ e $\bar{x}$ sono, rispettivamente, la media dei valori Y_i e x_i.

Un approccio alternativo, che porta allo stesso valore dei coefficienti β_0 e β_1 è il *principio dei minimi quadrati*.

Principio dei minimi quadrati. Detto *residuo i-esimo* la differenza verticale fra l'osservazione i-esima e la retta di regressione lineare:

$$E_i = Y_i - (\beta_0 + \beta_1 x_i),$$

e detta $f(\beta_0, \beta_1)$ la funzione somma dei quadrati dei residui:

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2 = \sum_{i=1}^{n} E_i^2$$

le stime $\hat{\beta}_0$ e $\hat{\beta}_1$ si ottengono minimizzando la funzione $f(\beta_0, \beta_1)$.

Esempio 6.23 *calcoliamo in R i coefficienti della retta di regressione lineare relativa all'esempio dei dati nel file* `cars`*:*

Soluzione.

```
> coefregr <- lm(dist~speed, data = cars)
> coef(coefregr)

##  (Intercept)        speed
##   -17.579095    3.932409
```

La retta ha quindi equazione:

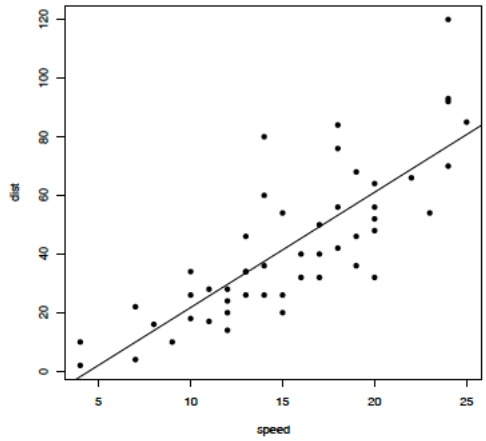
$$y(x) = -17.579095 + 3.932409x.$$

La prima parte dell'input della funzione `lm` è una formula che si legge: '`dist` è descritto da `speed`', che sono variabili all'interno del file specificato nel parametro `data`.

L'output `coefregr` contiene diverse informazioni, fra cui appunto i valori dei coefficienti della retta di regressione lineare.

È utile di solito visualizzare la retta insieme ai dati:

```
> plot(dist~speed, data = cars, pch = 16)
> abline(coef(coefregr))
```



Cerchiamo di capire il significato della retta di regressione e dei suoi parametri relativamente all'esempio precedente. La retta rappresenta l'andamento della distanza di arresto delle macchine rispetto alla loro velocità. Ovviamente aumentando la velocità aumenta la distanza di arresto, ma la retta indica *di quanto* aumenta la distanza all'aumentare della velocità. In particolare:

- Il coefficiente angolare β_1 indica la pendenza della retta. In questo esempio, rappresenta *di quanto* devo aumentare la distanza di arresto all'aumentare della velocità. Per ogni miglia oraria in più di velocità, devo aumentare di circa 3.93 piedi la distanza.
- L'intercetta β_0 è il valore in cui la retta intercetta l'asse y, quindi rappresenta la distanza di arresto di un'auto che ha una velocità pari a 0. Nella stima effettutata questa distanza è negativa (-17.58...), che ovviamente non ha senso. Guardando meglio i dati, notiamo che i valori della distanza sono presi a partire da una velocità di 4 miglia orarie, quindi stiamo cercando di ottenere un dato che è *all'esterno* dell'intervallo delle osservazioniu. Questa operazione viene detta *estrapolazione* ed è in pratica da evitare, perché può dare risultati senza senso, soprattutto se l'estrapolazione avviene lontano dagli estremi dell'intervallo dei dati.
- Quindi cosa significa $y = \beta_0 + \beta_1(8)$? Significa che se l'auto ha una velcoità di 8 miglia orarie, la distanza di arresto viene calcolata come $\beta_0 + \beta_1(8) = -17.58 + (8)(3.93) = 13.88$.

6.4.3 Predizione di valori futuri

Come abbiamo visto nell'esempio precedente, la retta di regressione lineare può essere utilizzata per *predire* valori futuri. Per fare questo, basta calcolare, nell'equazione della retta, il valore di y in corrispondenza della x desiderata. Si può fare manualmente, sostituendo i valori nell'equazione della retta, oppure in R con la funzione `predict`.

Nell'esempio precedente, abbiamo in realtà anche il valore della distanza corrispondente a 8 miglia orarie:

```
> cars[1:5, ]

##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
```

che sarebbe di 16 piedi. Possiamo interpretare il valore di 13.88 ottenuto come il valore di distanza di arresto più probabile per una futura macchina che va a 8 miglia orarie.

I valori $\hat{Y}_i$ ottenuti dalla retta di regressione lineare in corrispondenza delle ascisse x_i vengono detti *valori fittati*, mentre i valori ottenuti in corrispondenza di altre ascisse, diverse da $x_i, i = 1, \ldots n$, sono detti *valori predetti.*

Non è consigliabile predire un valore $\hat{Y}$ corrispondente ad un valore x al di fuori dell'intervallo delle osservazioni $[x_{min}, x_{max}]$.

Esempio 6.24 *Utilizziamo la funzione R `predict` per predire alcuni valori.*

Soluzione.

```
> regr_cars <- lm(dist~speed, data = cars)
> fitted(regr_cars)[1:5]

##         1         2         3         4         5
## -1.849460 -1.849460  9.947766  9.947766 13.880175

> predict(regr_cars,
+         newdata = data.frame(speed=c(7, 9, 2)))

##         1         2         3
##  9.947766 17.812584 61.069080
```

6.4.3.1 Coefficiente semplice di determinazione

È possibile avere un *singolo numero* che mi dà indicazioni sulla bontà del modello regressione lineare semplice rispetto al campione di dati a disposizione?

Il *coefficiente semplice di determinazione* viene calcolato appunto per questo scopo. Esso è definito dalla formula:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$

dove, ricordiamo:

- $Y_i, i = 1, \dots n$ sono i valori 'osservati' del campione;
- $\hat{Y}_i, i = 1, \dots n$ sono i valori 'fittati', cioè i valori del modello di regressione lineare semplice in corrispondenza delle ascisse x_i ($\hat{Y}_i = \beta_0 + \beta_1 x_i, i = 1, \dots n$);
- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ è la media dei valori osservati.

Si ha che $0 \leq r^2 \leq 1$. Tanto più r^2 è vicino a 1, tanto più il modello di regressione lineare è *buono*; tanto più r^2 è vicino a 0, tanto più il modello non è rappresentativo del campione dei dati. In quest'ultimo caso, l'analista cerca un modello differente da quello lineare per rappresentare i dati (una regressione non lineare o multivariata che coinvolga più di una variabile per esempio).

Associato al coefficiente semplice di determinazione r^2 si utilizza il *coefficiente semplice di correlazione* r che si ottiene come:

$$|r| = \sqrt{r^2}.$$

Per quanto riguarda il segno di r, si assume il segno della stima di β_1 calcolata.

Esempio 6.25 *Calcolare il valore del coefficiente r^2.*

Soluzione.

Il valore di r^2 è fra i valori riportati dal comando `summary`, sotto la dicitura: 'Multiple r squared'.

```
> regr_cars <- lm(dist~speed, data = cars)
> summary(regr_cars)

##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
```

```
## Signif. codes:
     0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511,Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Il valore $r^2 = 0.6511$ è basso. Infatti, il modello lineare in questo caso non rappresenta bene l'andamento delle osservazioni.

Il valore di r è:

```
> regr_cars <- lm(dist~speed, data = cars)
> carssum <- summary(regr_cars)
> sqrt(carssum$r.squared)

## [1] 0.8068949
```

Si sceglie il segno positivo perché il coefficiente angolare β_1 è positivo.

6.4.4 Inferenza sui parametri

I parametri $\hat{\beta}_0$ e $\hat{\beta}_1$ sono anch'essi variabili aleatorie e dipendono dal campione considerato. Su di essi, pertanto, si possono fare le inferenze di tipo statistico che abbiamo visto nei paragrafi precedenti, quali stime di intervalli di confidenza e test di ipotesi.

Per quanto riguarda la stima di intervalli di confidenza, per il parametro $\hat{\beta}_1$ possiamo dire che la sua distribuzione è normale con media uguale a β_1 e deviazione standard:

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\sum_i (x_i - \bar{x})^2}},$$

che dipende dalla deviazione standard esatta σ della distribuzione dei campioni, spesso non nota. Si utlizza quindi al posto di σ la sua stima S, ottenendo la seguente stima per la deviazione standard di $\hat{\beta}_1$:

$$S_{\hat{\beta}_1} = \frac{S}{\sqrt{\sum_i (x_i - \bar{x})^2}}.$$

Quindi, l'intervallo di confidenza per $\hat{\beta}_1$ è dato da:

$$\hat{\beta}_1 \pm t_{\alpha/2}(df = n-1) S_{\hat{\beta}_1}.$$

Analogamente si procede per calcolare l'intervallo di confidenza di $\hat{\beta}_0$.

Esempio 6.26 *Calcoliamo l'intervallo di confidenza per i parametri della retta di regressione lineare stimati in R.*

Soluzione.

Utilizzando la funzione `summary` di R:

```
> regr_cars <- lm(dist~speed, data = cars)
> summary(regr_cars)

##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:
     0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511,Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12

> confint(regr_cars)

##                 2.5 %    97.5 %
## (Intercept) -31.167850 -3.990340
## speed         3.096964  4.767853
```

Nella sezione `Coefficients` si hanno, oltre alle stime dei parametri calcolate dalla funzione `lm`, i rispettivi errori standard nella seconda colonna. Gli intervalli di confidenza possono essere calcolati sia richiamando la funzione R `confint` come nel codice (ricordiamo che il default è un livello di significatività $\alpha = 0.05$), sia con la formula precedente:

$$\hat{\beta}_1 \pm t_{0.025}(df = 23) S_{\hat{\beta}_1} = \hat{\beta}_1 \pm t_{0.025}(df = 23) 0.415.$$

Quindi, con il 95% di probabilità, il parametro β_1 sta nell'intervallo casuale $[3.097, 4, 768]$.

Per il test di verifica di ipotesi, il parametro più importante è sicuramente $\hat{\beta}_1$ rispetto a $\hat{\beta}_0$. Per $\hat{\beta}_1$ il test di ipotesi più frequente è quello che verifica se $\hat{\beta}_1 \neq 0$, cioè se la retta di regressione è parallela all'asse x oppure no. Una retta di regressione parallela all'asse x significa che il valore della variabile aleatoria Y NON cambia al variare della variabile indipendente x. Quindi il test di ipotesi è formulato come:

- H_0: $\beta_1 = 0$
- H_a: $\beta_1 \neq 0$

Esempio 6.27 Eseguiamo in R un test di ipotesi sul coefficiente angolare della retta di regressione lineare.

Soluzione.

In R, si utilizza la funzione `summary` nella sua sezione `coefficients` in cui vengono fornite informazione sui coefficienti.

```
> regr_cars <- lm(dist~speed, data = cars)
> summary(regr_cars)

##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:
     0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511,Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

In queste informazioni (ultima colonna) c'è anche il p-value per il test di ipotesi sia su $\hat{\beta}_1$ che su $\hat{\beta}_0$ (facendo su $\hat{\beta}_0$ lo stesso test che abbiamo fatto su $\hat{\beta}_1$). Ricordiamo che il p-value dà indicazioni sul livello di significatività del test di verifica. In particolare, è il più piccolo valore di significatività per cui l'ipotesi nulla viene rigettata. Quindi se il p-value è molto piccolo, è probabile che l'ipotesi nulla venga rigettata. Siccome in questo caso il p-value è minore di 0.05 (anche per β_1), possiamo dire che è praticamente certo che sia β_1 che β_0 siano non nulli.

6.4.5 Inferenza sui valori predetti

Rispondiamo ora a questa domanda:

' quanto sono *buoni* i valori predetti per una certa x_0? Quanta affidabilità ha questa stima? '

Anche i valori predetti sono variabili casuali che dipendono dal campione considerato.

Dobbiamo quindi calcolare un intervallo di confidenza per il valore predetto $\hat{Y}_0 = \hat{\beta}_1 x - i + \hat{\beta}_0$.

Per calcolare l'intervallo, dobbiamo partire dalla distribuzione della variabile aleatoria $\hat{Y}(x)$.

Nel caso dei valori 'predetti', la distribuzione di $\hat{Y}_{new}$ è normale con deviazione standard:

$$sd = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Poiché σ, deviazione standard della distribuzione da cui sono estratti i campioni, non è nota, si approssima con la deviazione standard dei campioni S. Quindi, seguendo le formule viste in precedenza l'intervallo di confidenza in questo caso è dato da:

$$\hat{Y}_{new} \pm t_{\alpha/2}(df = n-2)S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Esempio 6.28 *Calcolare in R l'intervallo di confidenza per alcuni valori predetti.*

Soluzione.

In R questi intervalli di confidenza si calcolano con la funzione `predict`, utilizzando l'argomento di input `interval`.

```
> regr_cars <- lm(dist~speed, data = cars)
> ndata <- data.frame(speed = c(5, 6, 21))
> predict(regr_cars, newdata = ndata,
+         interval = "predict")
```

```
##         fit       lwr      upr
## 1  2.082949 -30.33359 34.49948
## 2  6.015358 -26.18731 38.21803
## 3 65.001489  33.42257 96.58040
```

In questo caso di default l'intervallo di confidenza è al 95%. Se vogliamo cambiare il livello, si utilizza il parametro di input `level` nella funzione `predict`. Come si vede dall'output della funzione, gli intervalli di confidenza di predizione sono molto ampi.

6.4.6 Analisi dei residui

I residui per il modello di regressione lineare si possono ottenere in R con la funzione `residuals`:

Esempio 6.29 Visualizzazione dei residui in R.

Soluzione.

```
> regr_cars <- lm(dist~speed, data = cars)
> residuals(regr_cars)[1:5]
```

```
##         1         2         3         4         5
##  3.849460 11.849460 -5.947766 12.052234  2.119825
```

È importante stimare la varianza σ^2 dei residui, perché il parametro σ^2 determina la quantità di variazione nel modello di regressione lineare.

In particolare, se σ^2 è grande, i dati osservati (x_i, Y_i) saranno distanti dalla retta di regressione lineare, viceversa, se σ^2 è piccolo i dati saranno molto vicini alla retta.

Per stimare σ^2, utilizzando il metodo MLE, si ottiene il seguente stimatore::

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} E_i^2.$$

Per una stima puntuale si utilizza però di solito l' *errore quadratico medio* S^2 definito da:

$$E^2 = \frac{\sum_{i=1}^{n} E_i^2}{n-2},$$

dove E_i è l'i-esimo residuo, $E_i = Y_i - \hat{Y}_i$ dove $\hat{Y}_i$ è il *valore fittato di* Y_i.

Per stimare la deviazione standard uso l' *errore standard*

$$E = \sqrt{E^2}.$$

Esempio 6.30 *Uso del comando* `summary` *in R.*

Soluzione.

Con il comando `summary` fra le diverse informazioni abbiamo anche il valore del *residual standard error* che è appunto il valore E stima di σ.

```
> summary(regr_cars)

##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:
     0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511,Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Il modello di regressione lineare semplice è basato sull'ipotesi che i residui abbiano una distribuzione normale con media $\mu = 0$ e siano fra loro indipendenti.

In questo paragrafo ci occuperemo di verificare le ipotesi sui residui che stanno alla base dell'utilizzo del modello di regressione lineare semplice.

In particolare, come esemplificazione, consideriamo solo la verifica della normalità dei residui, lasciando ad un ulteriore approfondimento su altri testi le altre verifiche.

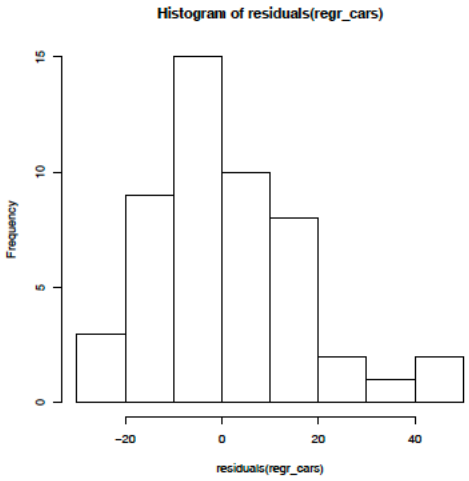
Possiamo testare la normalità dei residui con metodi grafici e test di ipotesi.

Esempio 6.31 Testiamo graficamente la normalità dei residui.

Soluzione.

Graficamente, possiamo fare un istogramma e un q-q-plot.

```
> regr_cars <- lm(dist~speed, data = cars)
> hist(residuals(regr_cars))
```

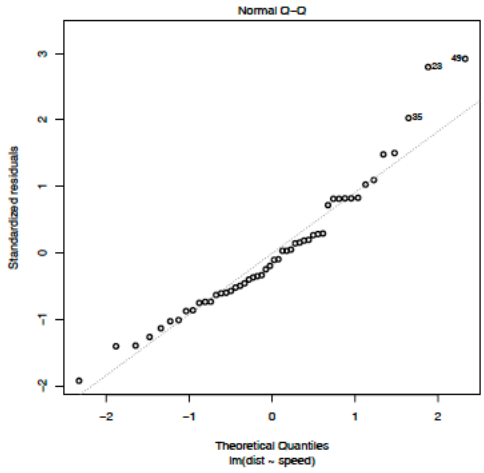



```
> plot(regr_cars, which = 2)
```

L'istogramma risulta non perfettamente simmetrico, e questo è un segno di non perfetta normalità.

Per quanto riguarda il grafico q-q plot, se la distribuzione dei residui fosse normale, dovrebbero essere casualmente distribuiti su entrambi i lati della retta tratteggiata.

In realtà, da un certo punto in poi, i residui sono tutti distribuiti dalla stessa parte della retta (anche se non lontani) e anche questo è un indice di non perfetta normalità.



Esempio 6.32 *Facciamo ora un test di verifica di ipotesi sulla normalità dei residui.*

Soluzione.

Ci sono diversi possibili test per questo tipo di verifica. Consideriamo il test di Shapiro-Wilk, basato sulla statistica:

$$SW = \frac{(\sum_{i=1}^{n} a_i E_{(i)})^2}{\sum_{j=1}^{n} E_j^2},$$

dove $E_{(i)}$ sono i residui ordinati, a_i sono delle costanti, $E_j = \hat{Y}_j - Y_j$ sono i residui.

Test di ipotesi:

- H_0: *i residui sono distribuiti in modo normale*
- H_a: *i residui NON sono distribuiti in modo normale*

Esempio 6.33 *Eseguiamo in R il test di ipotesi sulla normalità dei residui usando il test di Shapiro-Wilk.*

Soluzione.

In R, posso utilizzare la funzione `shapiro.test` del pacchetto *stats*:

```
> regr_cars <- lm(dist~speed, data = cars)
> shapiro.test(residuals(regr_cars))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(regr_cars)
## W = 0.9451, p-value = 0.02152
```

Quindi, per esempio, per un livello di significatività $\alpha = 0.05$ l'ipotesi di normalità viene rigettata.

È da ricordare comunque che il modello di regressione lineare semplice è *robusto* anche *sufficientemente lontano* dalla normalità dei residui. Questo significa che, anche se i residui non sono esattamente normali, il modello è comunque ragionevolmente valido e si può in pratica utilizzare.

Esercizi

Scrivere programmi R per:

Esercizio 6.1 *Considerare SRS(n) da una distribuzione normale con media e deviazione standard assegnate. Verificare che la distribuzione della media campionaria è* `norm(mean=mu, sd=`$\sigma/\sqrt{8n}$`)` *al variare del numero di elementi n del campione e del numero k di campioni considerati.*

Esercizio 6.2 *Verificare il teorema del Limite Centrale per SRS(n) da una distribuzione di probabilità che non sia n normale n lognormale.*

- *Visualizzare graficamente le distribuzioni ottenute per la media campionaria $\bar{X}$ al variare di n e confrontarle con la distribuzione normale che si dovrebbe ottenere.*
- *Visualizzare inoltre l'errore fra la media campionaria $\bar{X}$ simulata e quella esatta della distribuzione da cui sono stati estratti i campioni.*

Esercizio 6.3 *Stimare con il metodo MLE il parametro λ di una distribuzione di Poisson, simulando di avere campioni dalla distribuzione stessa. Calcolare il MSE al variare del campione.*

Esercizio 6.4 *Stimare con il metodo MLE il numero di gradi di libertà di una distribuzione chiquadro, simulando di avere campioni dalla distribuzione stessa.Calcolare il MSE al variare del campione.*

Esercizio 6.5 *Stimare con il metodo MLE parametri* `shape` *e* `rate` *di una distribuzione gamma, simulando di avere campioni dalla distribuzione stessa.Calcolare il MSE al variare del campione.*

Esercizio 6.6 *Calcolare l'intervallo di confidenza della media μ di una distribuzione normale* `norm(mu=20, sd=4)` *in cui è nota la deviazione standard, simulando un SRS(n) ndella distribuzione, al variare di n e del livello di confidenza $1-\alpha$. Visualizzare anche graficamente i risultati ottenuti, confrontandoli con la media esatta che si doveva ottenere.*

Esercizio 6.7 *Ripetere l'esercizio precedente nel caso di una distribuzione normale in cui non è nota la devizione standard (quindi utilizzando la deviazione standard campionaria nella formula).*

Esercizio 6.8 *Fare una simulazione di test di ipotesi sulla media di un campione estratto da una distribuzione normale con media $\mu = 2$ e deviazione standard $\sigma = 1.5$. Eseguire il test con diversi livelli di significatività α. Calcolare il p-value. (Considerare come ipotesi nulla sia il valore esatto della media $\mu = 2$ che un valore approssimato, per esempio $\mu = 2.01, 2.05$ ecc.)*

- *Ripetere la simulazione precedente supponendo di NON conoscere la deviazione standard σ.*

- *Ripetere la simulazione estraendo il campione da una distribuzione NON normale. Considerare un valore di n, numero degli elementi del campione, tale da poter applicare il Teorema del limite centrale.*

Esercizio 6.9 *Svolgere un'analisi di regressione lineare su un file di dati a piacere, eseguendo i seguenti step:*

- *scatter plot dei dati*
- *calcolo dei coefficienti della retta di regressione lineare*
- *plot della retta all'interno dello scatter plot*
- *calcolo del coefficiente di determinazione r^2*
- *analisi dei residui: plot dei residui, verifica della normalità dei residui con un test di ipotesi, verifica della media $\mu = 0$ della distribuzione normale.*
- *verifica del* non parallelismo rispetto all'asse x *della retta di regressione attraverso un test di ipotesi sul coefficiente angolare*