Data Mining
A.A. 2015-2016

GIACOMO MANZOLI
1130822
Università degli Studi di Padova
24 dicembre 2023

Indice

Capitolo 1

Introduzione al corso

Con l'evoluzione tecnologica abbiamo raggiunto un punto in cui costa poco rilevare dati in modo automatico e allo stesso tempo costa poco memorizzarli in data base, da qui il termine **big data**.

Datawarehouse: database che contiene tutti i dati su cui fare mining, i dati contenuti possono essere estratti da vari database che non sono necessariamente uniformi. Tipicamente non sono necessarie tutte le informazioni presenti nel DWH, pertanto è necessario tenere in considerazione gli obiettivi delle analisi in modo da utilizzare solamente i dati utili (**datamart**). Avendo a disposizione un elevato numero di campioni, può tornare utile estrarre un sottoinsieme casuale.

1.1 Struttura del corso

Parte 1:

- Modello lineare
- Inferenza sui parametri del modello
- Estensione del modello
- Modello lineare

Parte 2:

- Tecniche statistiche per l'approccio analitico
- Tecniche generali per la selezione di un modello
- Metodi di previsioni di variabili quantitativi
- Modelli di classificazioni (lineari e generalizzazione) (NN, Alberi, ecc.)
- Cenni ai metodi di raggruppamento (apprendimento non supervisionato)

Circa metà delle lezioni saranno in laboratorio, utilizzeremo vari software e la scelta è libera, a condizione che venga raggiunto l'obiettivo.

Come linguaggio di programmazione utilizzeremo **R**.

L'esame si concentra su:

- conoscenza delle metodologie studiate

- capacità di applicare

Prova parziale alla fine della prima parte con esercizi e domande di teoria. Al termine della seconda parte un progetto applicativo utilizzando gli strumenti appresi durante il corso (con relazione e discussione orale).

`homes.stat.unipd.it/bruno/dm-inf`
`scarpa@stat.unipd.it`

Capitolo 2

Introduzione al data mining

Il tutto è nato dalla presenza di data set enormi che hanno portato dei problemi sia a livello di gestione dei dati, che nell'analisi efficace dei dati.

L'elevata dimensionalità dei dati porta un problema sia computazionale, che concettuale, dal momento che risulta più difficile trovare le correlazioni tra i dati. Sono pertanto necessarie strutture dati più complesse come reti o alberi.

2.1 Data mining

Data mining: attività di elaborazione in forma grafica o numerica di grandi raccolte o di flussi continui di dati con lo scopo di estrarre informazione utile a chi detiene i dati stessi.

L'informazione utile può essere caratterizzata come:

- il comportamento complessivo del fenomeno in esame, con l'obiettivo di costruire un modello globale a partire dai dati disponibili
- l'individuazione di particolarità, ovvero delle caratteristiche specifiche di un piccolo insieme di dati.

Il data mining riguarda vari settori:

- l'informatica, per la gestione e il trattamento dei dati
- l'intelligenza artificiale, per la definizione di tecniche automatiche per il trattamento dell'informazione (machine learning)
- alla base di tutto c'è pero la statistica che viene applicata per trattare i dati.

2.1.1 Complicanze

- Enorme quantità e dimensioni dei dati
- I dati vengono raccolti per scopi diversi da quelli per cui vengono analizzati. Generalmente vengono per esigenze gestionali o simili e non per fini statistici (**ambito osservazionale**).
  - non esiste un **piano di campionamento** ma si ha a disposizione solo quello che si riesce ad osservare

  - i dati sono sporchi
  - molto spesso ci sono dei dati mancanti

- spesso si va alla ricerca di qualcosa, senza sapere che cosa e si trova comunque qualcosa.

Capitolo 3

Modello lineare semplice

- **OLTP**: strumenti di interrogazione su specifiche informazioni da rihicedere ai vari database, detti operativi
- **OLAP**: ...
- **KDD**: Knowledge discovery database, si parte da uno o più database operativi per costruirne uno strategico, il data whare house; questa costruzione comporta anche un'operazione di omogeneizzazione di definizione di variabili e operazioni di pulizia dei dati (**data mining analitico**).

Modello: (o algoritmo) rappresentazione semplificata del fenomeno di interesse, funzionale ad un obiettivo specifico.

Non esiste un modello vero in quanto si tratta di approssimazioni molto dettagliate, ci sono dei modelli che in determinati contesti risultano migliori di altri. Specialmente in ambiti non scientifici, il criterio per la bontà di un modello è il **basta che funzioni**, questo perché tipicamente i dati che vengono utilizzati non sono stati raccolti con un criterio sperimentale.

Tipicamente il modello viene visto come una scatola nera, funziona ma non si sa quale sia il vero meccanismo che regola il fenomeno. Tuttavia, questa black box deve essere comunque manutenuta, non basta avere solamente l'hardware e il software.

3.1 Il modello lineare semplice

Si parte da due variabili e si costruisce un modello che li mette in relazione tra loro.

Il data set di riferimento riguarda i dati di 200 mercati su cui opera un'azienda, per i quali si conosce la quantità di merce venduta in migliaia e il budget speso per la pubblicità radiofonica in quella zona.

Si vuole ottenere un'equazione che permetta di prevedere le vendite in funzione del budget.

Il primo passo è quello del costruire un grafico di dispersione, dove nelle x ci sono le spese e nelle y ci sono le unità vendute.

Dal grafico si può osservare un andamento lineare che può essere approssimata con:

$$vendite = \beta_0 + \beta_1(radio) + (errore)$$

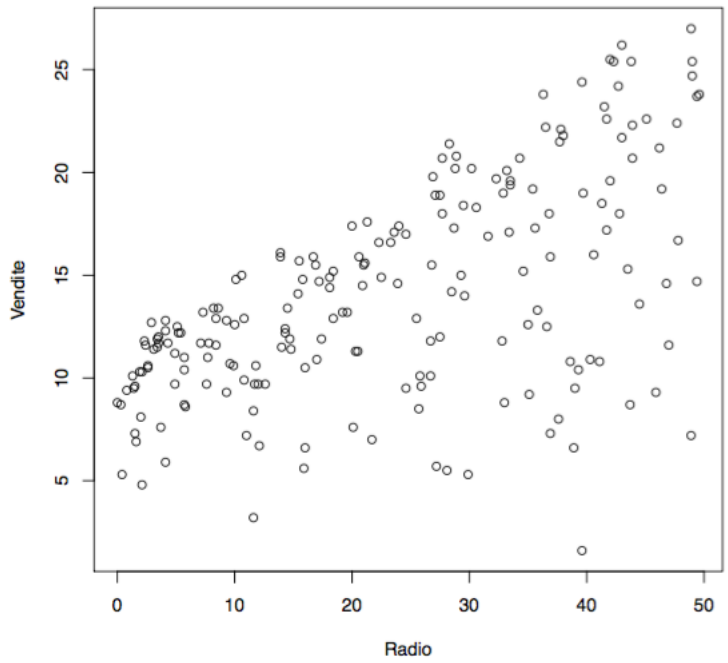


Figura 3.1: Grafico di dispersione per il data set delle vendite

dove la componente *errore* esprime la parte delle vendite non legate alle pubblicità via radio.

Un modello di questo tipo prende il nome di **modello di regressione lineare semplice**.

La variabile y (*vendite*) prende il nome di **variabile risposta/dipendente/output** mentre x (*radio*) prende il nome di **variabile esplicativa/indipendente/input** e i vari coefficiente β prendono il nome di **parametri**.

Il gioco adesso diventa quello di andare a trovare dei valori $\hat{\beta}_0$ e $\hat{\beta}_1$ che approssimano la retta nel miglior modo possibile.

La ricerca avviene utilizzando i dati presenti nel data set:

$$
\begin{aligned}
y_1 &\approx \hat{\beta}_0 + \hat{\beta}_1 x_1 \\
y_2 &\approx \hat{\beta}_0 + \hat{\beta}_1 x_2 \\
&\dots \\
y_n &\approx \hat{\beta}_0 + \hat{\beta}_1 x_n
\end{aligned}
$$

Raffinando l'idea si ottiene il metodo dei **minimi quadrati**

$$s^2(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

e si vanno a cercare i parametri che minimizzano l'errore di stima ai minimi quadrati.

$$s^2(\hat{\beta}_0, \hat{\beta}_1) \leq s^2(\beta_0, \beta_1)$$

Si utilizza il quadrato della distanza, sia per rendere l'errore indipendente dal segno, sia per dare maggior peso ad errori maggiori.

Ci sono un po' di barbatrucchi matematici per trovare il minimo quello che interessa è:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Da notare che nel lato pratico non ci sarà mai un dataset con varianza nulla, perché in quel caso il problema di regressione non ha senso. Per il dataset delle vendite si ottiene come retta ai minimi quadrati

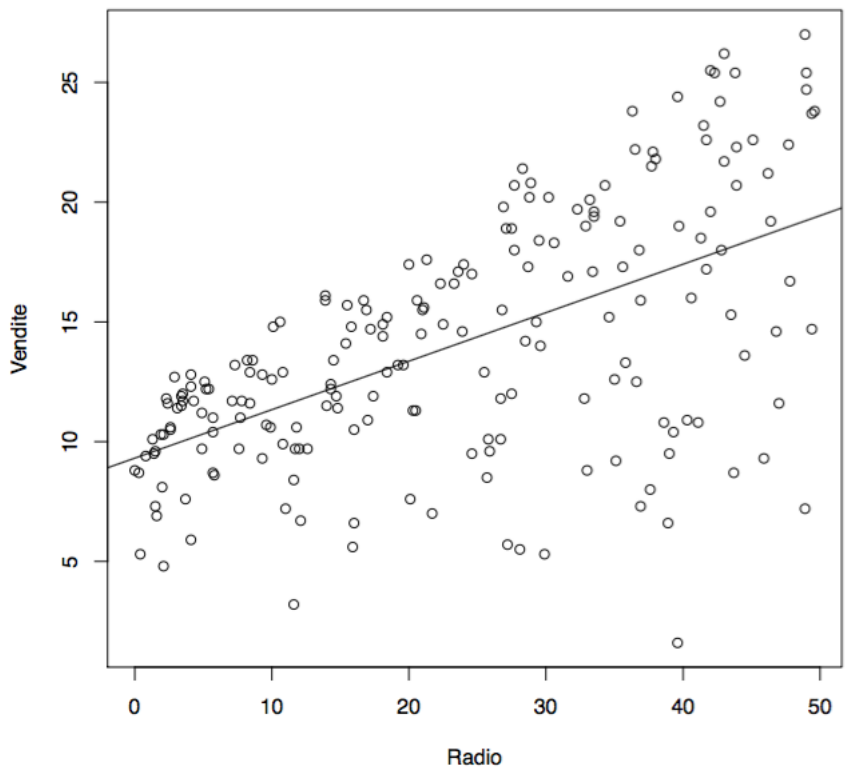


Figura 3.2: Retta ai minimi quadrati per il dataset delle vendite

3.1.1 Residui

Non è detto che la retta ai minimi quadrati approssimi bene i dati.

Un indicatore dell'andamento è dato dai **residui**, che rappresentano la differenza tra i valori osservati e quelli ottenuti utilizzando la retta.

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \ \forall i = 1, \ldots, n$$

Per costruzione della retta, la somma di tutti i residui risulta essere 0:

$$\begin{aligned}\sum_{i=1}^{n} r_i &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\ &= n\bar{y} - n(\bar{y} - \hat{\beta}_1 \bar{x}) - n\hat{\beta}_1 \bar{x} \\ &= 0\end{aligned}$$

Quindi per valutare la retta è necessario utilizzare la **varianza** dei residui. Minore è la varianza, migliore è la retta.

$$\begin{aligned}\mathrm{var}(r_1, \ldots, r_n) &= \frac{1}{n}\sum_{i=1}^{n}(r_i - 0)^2 \\ &= \mathrm{var}(Y) - \frac{\mathrm{cov}^2(X, Y)}{\mathrm{var}(X)}\end{aligned}$$

Nel caso peggiore, la varianza dei residui ha come bound superiore la varianza della risposta, ovvero $\mathrm{var}(Y)$.

Più piccola è la varianza dei residui, più la retta di regressione "spiega" le variazioni della risposta.

È quindi possibile definire il **coefficiente di determinazione**:

$$R^2 = 1 - \frac{\mathrm{var}(r_i, \ldots, r_n)}{\mathrm{var}(Y)}$$

R^2 varia tra 0 e 1, dove 1 è il valore ottimo e 0 è il valore peggiore.

Da notare che ciò vale solo per la retta.

3.2 Valutazione del modello lineare semplice

R^2 è un numero compreso tra 0 e 1, ma quanto bene va il mio modello nel predirre il futuro?

Il modello ottenuto prende in considerazione i dati sui quali è stato calcolato ed è corretto assumere che, se vengono replicati gli stessi dati, ci sarà comunque un fattore di diversità tra la stime prodotte e i dati osservati.

La domanda diventa quindi, quanto i coefficienti stimati sono diversi dai valori corretti?

Ad esempio, se si ottiene un $\hat{\beta}_1$ uguale a 0 non si può dire che le x e le y sono tra loro indipendenti, ma che il modello lineare non è utile per spiegare y utilizzando x.

Quindi, quanto è l'errore che otteniamo con il modello ipotizzato? Quando possiamo dire che R^2 è un valore sufficientemente buono?

Tornando al modello delle vendite, queste sono composte da due componenti:

- una componente **sistematica** (deterministica) che viene determinata dal modello di regressione e che dipende dalla spesa in pubblicità

- una componente **aleatoria** o erratica, che varia in modo stocastico e rappresenta le vendite dovute al caso e non prevedibili.

La componente erratica può essere modellata utilizzando una variabile aleatoria ϵ con media nulla, il che vuol dire che le previsioni che risultano maggiori del valore medio sono confrontabili con quelle che risultano minori.

È inoltre ragionevole assumere che ϵ sia indipendente sia da x che da y, ovvero che tutta la relazione tra x e y viene vincolata dalla componente sistematica. Si può quindi dire che la varianza di ϵ è costante.

Le ipotesi sul modello diventano quindi

- e_i hanno tutte la stessa distribuzione
- $E(e_i) = 0 \, \forall i$
- per ogni $i \neq j$, e_i è indipendente da e_j
- Le varianze degli e_i sono costanti.

Con questa definizione si ha che anche y è una variabile aleatoria di media $\hat{\beta}_0 + \hat{\beta}_1 x$ e varianza σ^2, questo perché è una trasformazione lineare di una variabile aleatoria:

dalla teoria delle probabilità si ha che se Z è una variabile aleatoria e a e b sono due costanti, allora:

$$E(aZ + b) = aE(Z) + b$$

e

$$\text{var}(aZ + b) = a^2 \text{var}(Z) + b$$

Sempre per quanto riguarda la distribuzione degli errori, sembra essere ragionevole che la questa sia simmetrica rispetto alla media e che sia più probabile sbagliare poco, ovvero che segua la distribuzione gaussiana.

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Pertanto anche y diventa una variabile aleatoria gaussiana.

$$y \sim \mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x, \sigma^2)$$

3.2.1 Valutazione dell'errore di stima su $\hat{\beta}_1$

Riassumendo:

$$y = \beta_0 + \beta_1 x + \epsilon \qquad y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$

La varianza degli errori σ^2 viene assunta nota, al fine di semplificare i conti.

La stima di β_1 può quindi essere vista come una determinazione di una variabile casuale. Infatti, se le y da cui è calcolata sono il risultato di un esperimento casuale, anche $\hat{\beta}_1$ lo è:

$$\hat{\beta}_1 = \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x}^2)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x}^2)}$$

Quindi, se le n variabili casuali normali y_i hanno tutte la stessa varianza σ^2 si ha che

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1\,,\ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Questo perché la combinazione lineare di variabili casuali normali è anche essa una variabile casuale normale

$$a_0 + \sum_{i=1}^{n} a_i Z_i \sim \mathcal{N}\left(a_0 + \sum_{i=1}^{n} a_i \mu_i\,,\ \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Da ciò si può osservare che la varianza della stima di $\hat{\beta}_1$ coincide con la varianza delle stima delle y_i, scalata per una misura della dispersione delle x e che la media di $\hat{\beta}_1$ è β_1, ovvero la stima $\hat{\beta}_1$ è centrata sul valore vero del parametro, anche se questo è ignoto. Quando questo succede, lo stimatore si dice **non distorto** o **corretto**.

Dal momento che la varianza dello stimatore è influenzata dalla varianza dei campioni utilizzati, conviene cercare di diversificare il più possibile i campioni al fine di avere uno stimatore migliore.

Utilizzando il risultato precedente si può calcolare la distribuzione dell'**errore di stima**, ovvero $\hat{\beta}_1 - \beta_1$, che essendo una trasformazione lineare di una variabile aleatoria, risulta essere:

verifcare se l'affermazione è corretta

$$\hat{\beta}_1 - \beta_1 \sim \mathcal{N}\left(0\,,\ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Utilizzando questa nuova variabile aleatoria è possibile andare a calcolare la probabilità che l'errore di stima sia limitato in un certo intervallo, andando prima a trasformare la distribuzione nella normale $\mathcal{N}(0,1)$ e poi con le tabelle di conversione, andare a calcolare la probabilità effettiva.

$$\begin{aligned}
P(|\hat{\beta}_1 - \beta_1| < 0.05) &= P\left(\left|\mathcal{N}\left(0\,,\ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right| < 0.05\right) \\
&= P\left(\left|\frac{\mathcal{N}\left(0\,,\ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}\right| < \frac{0.05}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}\right) \\
&= P(|\mathcal{N}(0,1)| < 2.342) \text{ (valore recuperato dall'esempio delle slide)} \\
&= \Phi(2.342) - \Phi(-2.342) \approx 0.9808
\end{aligned}$$

In generale un intervallo che contiene il vero valore di un parametro ignoto con probabilità $1 - \alpha$ viene chiamato **intervallo di confidenza** $1 - \alpha$ e costituisce il modo più semplice di rappresentare la precisione di una stima. Tipicamente come valori di $1 - \alpha$ vengono utilizzati $0.5, 0.9, 0.95, 0.99$.

Più precisamente il procedimento per calcolare un intervallo di confidenza è:

1. Fissa un valore per $1-\alpha$.
2. Determina il percentile $1-\alpha/2$ della normale, ovvero $P(\mathcal{N}(0,1) \leq z_{1-\alpha/2}) = 1-\alpha/2$.
3. Si ha quindi che

$$P\left(\left|(\hat{\beta}_1 - \beta_1) \Big/ \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right| \leq z_{1-\alpha/2}\right) = 1-\alpha$$

4. L'intervallo di confidenza di livello $1-\alpha$ risulta quindi essere:

$$\left[\hat{\beta}_1 - \frac{z_{1-\alpha/2} \cdot \sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\,,\ \hat{\beta}_1 + \frac{z_{1-\alpha/2} \cdot \sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\right]$$

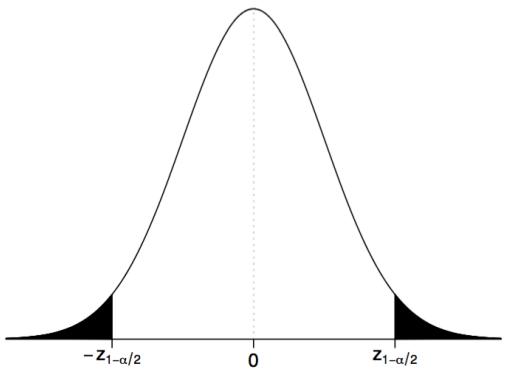


Figura 3.3: L'area in bianco equivale a $1-\alpha$

Verifica dell'ipotesi

Un modo alternativo di affrontare la valutazione dell'errore commesso stimando $\hat{\beta}_1$ è quello di chiedersi se il modello

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

può essere sostituito dal modello

$$y = \beta_0 + \epsilon$$

Si può quindi definire un problema di **verifica di ipotesi** utilizzando come **ipotesi nulla** e **ipotesi alternativa**, rispettivamente:

$$\begin{cases} H_0 : \beta_1 = 0 \rightarrow \text{ La spesa in pubblicità influenza le vendite} \\ H_1 : \beta_1 \neq 0 \rightarrow \text{ La spesa in pubblicità non influenza le vendite} \end{cases}$$

La scelta dell'ipotesi viene effettuata con un **test statistico** del tipo

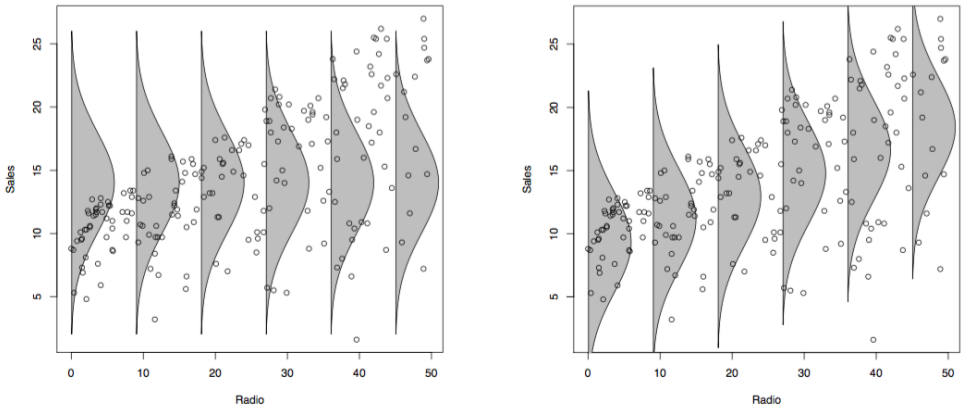


Figura 3.4: A sinistra H_0 e a destra H_1 per il problema delle vendite.

$$-h \leq \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \leq h$$

Resta da decidere un valore di soglia per h in modo che la probabilità di accettare H_0 quanto è vera sia 1, ovvero

$$P\left(-h \leq \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \leq h \text{ quando } \beta_1 = 0 \right) = 1$$

ma questo equivale a dire che motivare

$$P(-h \leq \mathcal{N}(0,1) \leq h) = 1$$

e questo è vero solo se $h = +\infty$. Questo valore di soglia però è inutile, perché si ha che H_0 viene accettata anche quando è falsa, pertanto ha più senso fissare una soglia che fornisca una probabilità sufficientemente alta, ovvero:

$$P(\text{accettare } H_0 \text{ quando è vera}) = 1 - \alpha$$

La formulazione diventa quindi

$$P(-h \leq \mathcal{N}(0,1) \leq h) = 1 - \alpha$$

ottenendo $h = z_{1-\alpha/2}$, dove con z_p viene indicato il percentile p-esimo della normale, ovvero il numero per cui $\Phi(z_p) = p$.

La procedura per accettare o rifiutare l'ipotesi nulla è quella riportata in figura **??**

Errori possibili

Con la formulazione delle due ipotesi è possibile commettere due tipi di errori:

- **errore di primo tipo**: viene rifiutata H_0 quando questa è vera.
- **errore di secondo tipo**: viene accettata H_0 quando H_0 è falsa.

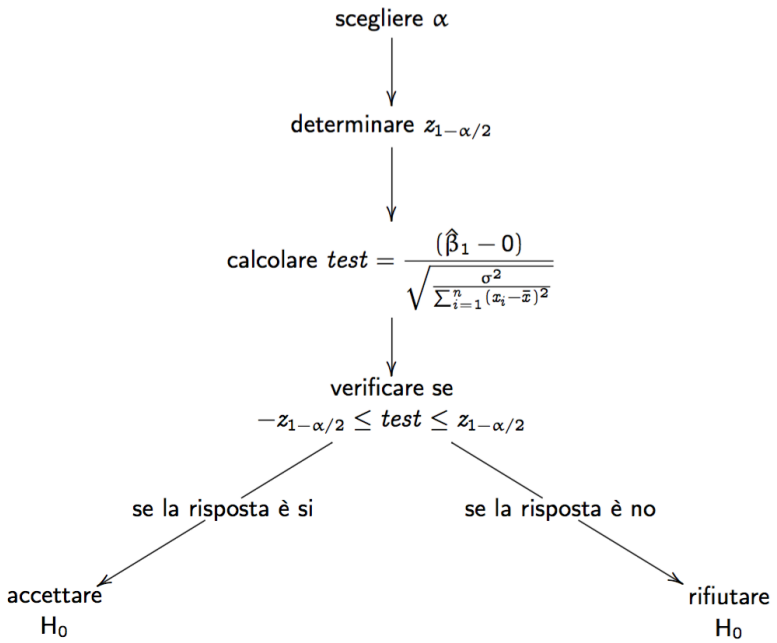


Figura 3.5: Procedure per l'accettazione o il rifiuto di H_0

Per come è stato costruito il test statistico si ha che

$$P(\text{errore } 1^\circ \text{ tipo}) = 1 - P(\text{accettare } H_0 \text{ quando è vera})$$

e viene fissato ad α, perché

$$P(\text{accettare } H_0 \text{ quando è vera}) = 1 - \alpha$$

Tuttavia non viene detto nulla a riguardo dell'errore del secondo tipo. Si sa solo che non si possono minimizzare i due errori contemporaneamente, perché sono tra loro correlati.

Ma alla fine quello che conta è l'errore di primo tipo, perché la domanda che ci si pone è "*A:i dati sperimentali sono compatibili con H_0?*" piuttosto che "*B:Quale tra H_0 e H_1 è vera?*".

L'approccio adottato prevede quindi di fissare l'errore più importante (con α) e minimizzare la probabilità d'errore del secondo tipo.

Sfruttando i dati è comunque possibile ridurre entrambi i tipi d'errore perché se il numero n di dati a disposizione è grande e quanto più è dispersa la variabile esplicativa, tanto più gli errori commessi sono minori.

3.2.2 Rilassando le ipotesi

Finora è stato assunto che i dati seguano una distribuzione normale, quindi prima di effettuare una regressione lineare è necessario andare a verificare che questo sia vero.

Per effettuare questa verifica servono delle tecniche grafiche ed analitiche complesse, pertanto è più ragionevole utilizzare un approccio alternativo che verifica a posteriori la normalità dei dati.

Un'altra ipotesi fatta riguarda la conoscenza della varianze delle distribuzione, ma nel lato pratico questo è un parametro ignoto.

Utilizzando il metodo dei minimi quadrati è possibile solamente stimare i coefficienti β e non la varianza.

Tuttavia, considerando $y = \beta_0 + \beta_1 x + \epsilon$, si può isolare ϵ e stimarlo utilizzando i residui osservati:

$$\begin{aligned}\hat{\epsilon}_i &= r_i \\ &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\end{aligned}$$

Si può quindi costruire uno stimatore per la varianza della componente di errore utilizzando la varianza dei residui[1] che è facilmente calcolabile.

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(r_i - 0)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\end{aligned}$$

Utilizzando questo stimatore si tende ad ottenere **una sottostima** della varianza, pertanto conviene utilizzare

$$\begin{aligned}\hat{s}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(r_i - 0)^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\end{aligned}$$

Il problema di verifica delle ipotesi senza σ^2

Dal momento che non è nota la varianza, non è più possibile utilizzare z per effettuare la verifica dell'ipotesi su β_1, ma è necessario utilizzare un'altra statistica test:

$$t_{oss} = \frac{(\hat{\beta}_1 - 0)}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

Il pedice *oss* indica che viene influenzata dai dati osservati.

Il funzionamento di t_{oss} è analogo a quello di z: se H_0 è vera ci si aspetta che t_{oss} assuma valori vicini allo 0, mentre se è vera H_1 ci aspettiamo che assuma valori lontani dallo 0.

Per determinare il valore di t_{oss} non è possibile utilizzare la distribuzione normale, perché la varianza non è nota, ed è quindi necessario utilizzare la ***t* di Student**[2].

La distribuzione *t di Student* dipende da un solo parametro che prende il nome di **gradi di libertà**. Durante a verifica dell'ipotesi sul coefficiente angolare della retta di regressione, il grado di libertà è $n - 2$, ovvero è stato dimostrato che:

[1] La media dei residui è 0.

[2] Questo è possibile perché nelle nostre ipotesi gli errori seguono la distribuzione normale e sono tra loro indipendenti.

$$t_{oss} \approx t \text{ di Student con } n-2 \text{ gradi di libertà.}$$

Il grafico di questa distribuzione è simile a quello della normale, solo che ha delle code più pesanti.

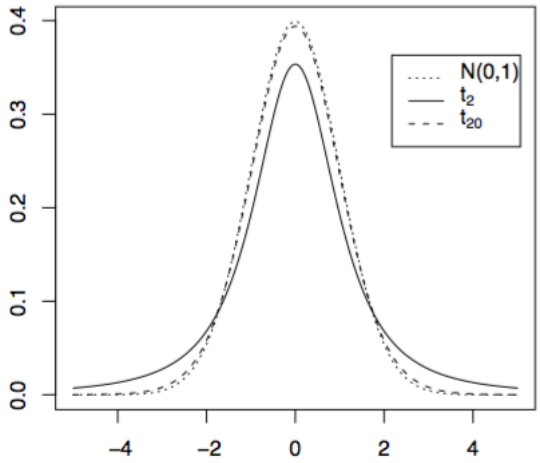


Figura 3.6: Confronto della t di Student con la normale $\mathcal{N}(0,1)$

In questo caso però non conviene utilizzare un confronto sulla soglia dei valori, perché questo potrebbe enfatizzare troppo delle differenze minime, portando a dei valori completamente diversi.

L'idea è quindi quello di utilizzare il **livello di significatività osservato** o **p-value**, un indicatore uguale alla probabilità di osservare, supponendo vera H_0, un valore di t_{oss} uguale o "più lontano" da H_0 di quanto effettivamente è osservato, ovvero costituisce una misura di quanto l'ipotesi nulla è plausibile sulla base dei dati osservati.

Cosa **NON** è il *p-value*:

- **Non** è la probabilità che l'ipotesi nulla sia vera o falsa. Non è connessa a nessuna delle due.
- **Non** è la probabilità che un'osservazione sia un caso.
- **Non** è la probabilità di rifiutare H_0 quando questa è vera
- **Non** è la probabilità di ottenere lo stesso risultato replicando l'esperimento.

Questo livello varia da 0 a 1 e più è grande, più i dati sono vicini ad H_0.

- Se vale 0 (o è inferiore ad α) vuol dire che sotto H_0 non è possibile osservare nessun altro valore più lontano da H_0, ovvero il valore osservato per t_{oss} è uno dei più lontani possibili. L'evidenza empirica è fortemente contraria all'ipotesi nulla che quindi va rifiutata.
- Se vale 1 (o è maggiore di α) vuol dire che sotto H_0 tutti i possibili valori osservabili per t_{oss} sono "non più vicini di quello osservato", ovvero quello osservato è uno dei più vicini possibili. L'evidenza empirica non è sufficientemente contraria all'ipotesi nulla, la quale non può essere rifiutata.

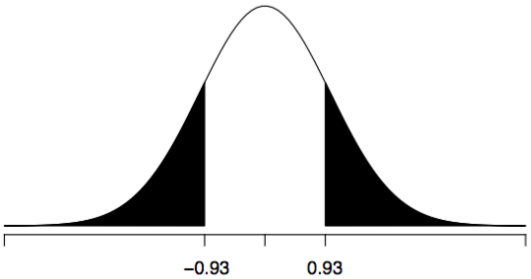
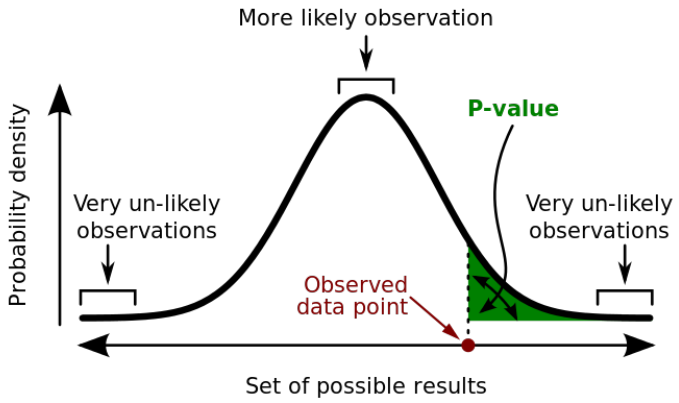


Figura 3.7: La curva mostra la densità di una t di Student con $t_{oss} = 0.93$. Poiché lontano da 0 vuol dire lontano da H_0, l'area annerita fornisce una approssimazione della probabilità di osservare, quando è vera H_0, un valore più lontano dall'ipotesi nulla rispetto a quanto osservato.

lontano da H_0 equivale a dire *"lontano da 0 in entrambe le direzioni"* quindi, nel caso delle vendite ($t_{oss} = 9.921$):

$$\begin{aligned}\text{livello di significatività osserato} &= P(|t \text{ con 198 gradi}| \geq t_{oss}) \\ &= P(|t \text{ con 198 gradi}| \geq 9.921) \\ &= 2xP(t \text{ con 198 gradi} \geq 9.921)\end{aligned}$$



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Il calcolo della probabilità viene poi fatto utilizzando una tabella o un software apposito. Nell'esempio si ha che il livello di significatività osservato è 0.

Ovvero, se la spesa in pubblicità via radio ha un effetto sulle vendite allora noi ci aspetteremo valori "Più lontani da H_0 di quanto osservato".

Il che vuol dire che il valore osservato di t_{oss} è molto strano (poco plausibile) quando H_0 è vera.

Interpretazione alternativa del *p-value* Supponendo di formare tutti i possibili campioni di 200 elementi e di calcolare t_{oss} per ognuno di questi. Allora il *p-value* è la percentuale

di valori di t_{oss} maggiori di 9.921[3], che per quanto calcolato prima ci si aspetta che sia 0. Ma allora 9.921 è un valore che difficilmente può capitare quando H_0 è vera, e quindi i dati smentiscono l'ipotesi nulla, che è da rifiutare.

La procedura per accettare o rifiutare l'ipotesi è riportata in Figura **??**.

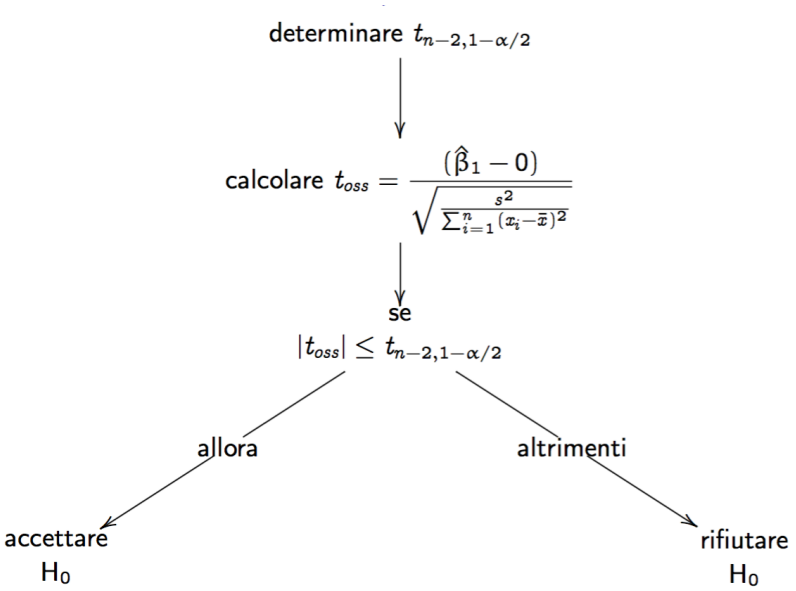


Figura 3.8: Regola per l'accettazione o il rifiuto di H_0. $t_{g,p}$ indica il percentile p-esimo di una t di Student con g gradi di libertà. L'albero fornisce una regola che garantisce di accettare H_0 quando questa è vera con probabilità $1 - \alpha$.

Intervallo di confidenza

In modo analogo a come è stato fatto nel caso di σ^2 noto è possibile definire un'intervallo di confidenza per la stima.

$$P\left(-t_{n-2,1-\alpha/2} \leq \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\frac{s^2}{\sum_{i=i}^{n}(x_i-\bar{x})^2}}} \leq t_{n-2,1-\alpha/2} \right) = 1 - \alpha$$

$$\left[\hat{\beta}_1 - t_{n-2,1-\alpha/2}\sqrt{\frac{s^2}{\sum_{i=i}^{n}(x_i-\bar{x})^2}}\,,\ \hat{\beta}_1 + t_{n-2,1-\alpha/2}\sqrt{\frac{s^2}{\sum_{i=i}^{n}(x_i-\bar{x})^2}} \right]$$

3.2.3 Verifica di ipotesi su β_0

In modo simile è possibile stimare β_0, anche se risulta molto meno interessante, perché è β_1 che fornisce la relazione tra le x e le y.

Come prima cosa è necessario stimare la varianza di $\hat{\beta}_0$:

$$\text{var}(\hat{\beta}_0) = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} s^2$$

Le ipotesi diventano

[3]Valore di t_{oss} calcolato per il data set d'esempio delle slide

$$\begin{cases} H_0 : \beta_0 = \beta' \\ H_1 : \beta_0 \neq \beta' \end{cases}$$

dove β' è un valore fissato, ad esempio 0.

La statistica test diventa quindi:

$$t_{oss} = (\hat{\beta}_0 - \beta') \Big/ s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

con intervallo di confidenza definito da

$$\left[\hat{\beta}_0 - t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}\ ,\ \hat{\beta}_0 + t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

Da notare che in questo modo non viene definito un intervallo di confidenza simultaneo, ovvero che vale sia per β_0 che per β_1.

3.2.4 Verifica di ipotesi globale sul modello

Per misurare la varianza spiegata dal modello abbiamo il coefficiente di determinazione R^2 che misura la varianza spiegata dal modello rispetto a quella complessiva. Non abbiamo però un criterio per capire se l'R^2 osservato è grande o piccolo.

Anche in questo caso è possibile impostare un problema di verifica di ipotesi, l'ipotesi nulla sarà l'indipendenza lineare tra x e y, mentre l'ipotesi alternativa prevederà che la retta spieghi, almeno in parte, la relazione presente.

Generalmente la statistica test più utilizzata per questo problema è:

$$F_{oss} = \frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}} = \left(\frac{R^2}{1-R^2}\right)\left(\frac{n-k}{k-1}\right)$$

dove k rappresenta il numero di parametri utilizzati nel modello, nel nostro caso 2.

La statistica F_{oss} può essere anche vista come il rapporto tra la media dei quadrati degli scarti spiegati dal modello e la corrispondente media dei quadrati degli scarti residui, pertanto ci aspettiamo F_{oss} grande quando H_0 è falsa.

Nelle ipotesi di normalità ed indipendenza degli errori, si ha che F_{oss} si distribuisce come una variabile aleatoria detta ***F* di Snedecor** con $k-1$ gradi di libertà al numeratore e $n-k$ gradi di libertà al denominatore.

Il valore F_{oss} deve essere quindi confrontato con la F di Snedecor utilizzando una tavola dei percentili o con un apposito software. Maggiore è F_{oss}, minore è la probabilità che H_0 sia vera.

Ad esempio, con $F_{oss} = 98.42$ si ha che il valore osservato è molto più grande del quantile 0.999(che vale circa 11.15), sia ha quindi che la probabilità di osservare un valore uguale o più lontano da H_0 rispetto a quando osservato è inferiore di un millesimo. I dati suggeriscono quindi che nel campione osservato si presenta una dipendenza lineare tra x e y e anche che questa dipendenza dovrebbe esserci in tutta la popolazione.

Il livello di significatività osservato fornisce quindi la probabilità di ottene una statistica F maggiore o uguale a quella ottenuta dai dati a disposizione nel caso ci sia indipendenza tra x e y.

Il vantaggio di questo livello è che nasconde tutti i dettagli riguardo la statistica test.

Tipicamente, se il livello di significatività osservato è inferiore a 0.01 i risultati sono considerati altamente significativi **contro** H_0, mentre se risulta compreso tra 0.01 e 0.05 i risultati sono comunque significativi contro H_0, ovvero è molto probabile che H_0 sia falsa.

Viceversa, se risulta maggiore di 0.1 si conclude che i dati non contengono elementi tali da poter rifiutare H_0 e quindi sono **non significativi**.

3.3 THE BIGGER PICTURE

un modello di regressione

come verificare queste evidenze *visive*?

- un modello di regressione semplice potrebbe essere:

$$\text{SAT} = \alpha + \beta \text{SPESA} + \varepsilon$$

con ε un errore assunto $WN(0, \sigma^2)$. La stima ai minimi quadrati fornisce

$$\hat{\alpha} = 1089.294 \ (\text{se} = 44.390) \qquad \hat{\beta} = -20.892 \ (\text{se} = 7.328)$$

- si può effettuare un test di significatività per verificare se β è significativamente diverso da zero.
- nelle usuali ipotesi la statistica del test sarà $t^{oss} = -2.85$ ($t_{48;0.975} = 2.01$, si rifiuta quindi l'ipotesi di uguaglianza a zero) e il livello di significatività osservato (*p-value*) sarà di $p = 0.0064$, che indica la forte significatività del parametro

  p-value è la probabilità di osservare un valore maggiore di t_oss nella t di student
- la relazione appare parecchio forte: ogni dollaro di aumento nella spesa per studente all'anno è associato con un *calo* di circa 21 punti nel punteggio medio SAT dello stato, e tale stima raggiunge ogni convenzionale livello di significatività statistica ($p < 0.01$)
- è possibile valutare la capacità esplicativa del modello ad esempio attraverso il calcolo dell'R^2 che risulta $R^2 = 0.144808$. Il test F sull'adattabilità complessiva del modello risulta pari a $F^{oss} = 8.13$ ($F_{1;47;0.975} = 5.36$) con *p-value* pari a $p = 0.0064$.

3.4 Effettuare previsioni

Una volta fatto il modello, è carino utilizzarlo per effettuare delle previsioni o delle simulazioni sul futuro.

Questo può essere fatto fissando un determinato valore x_0 e calcolando y utilizzando i coefficienti stimati:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Anche se il valore x_0 non è compreso nell'intervallo dei campioni è possibile effettuare una previsione (**estrapolazione**).

Tuttavia può essere interessante avere anche un intervallo di precisione da associare alla stima $\hat{y}_0$.

Partendo dal modello di riferimento si ottiene:

$$\begin{aligned}\hat{y}_0 &= \hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon \\ &= (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_0 + \epsilon \\ &= \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}) + \epsilon\end{aligned}$$

Si ha quindi che:

$$\begin{aligned}\mathrm{var}(\hat{y}_0) &= \mathrm{var}(\bar{y}) + \mathrm{var}(\hat{\beta}_1)(x_0 - \bar{x})^2 + \mathrm{var}(\epsilon) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)\end{aligned}$$

che può essere stimata utilizzando s^2 al posto di σ^2

$$\widehat{\mathrm{var}(\hat{y}_0)} = s^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

Da notare che all'aumentare del numero di campioni e all'aumentare della diversità delle x, la varianza della stima diminuisce.

L'intervallo di precisione risulta quindi essere

$$\left[\hat{y}_0 - t_{n-2,1-\alpha/2} \sqrt{\widehat{\mathrm{var}(\hat{y}_0)}} \, , \, \hat{y}_0 + t_{n-2,1-\alpha/2} \sqrt{\widehat{\mathrm{var}(\hat{y}_0)}} \right]$$

Nell'esempio delle vendite, se $x_0 = 60$ si ottiene $y_0 = 21.4614$ con un intervallo di previsione $[12.88 \, , \, 30.04]$ valido con il 95% di probabilità.

Effettuando il calcolo dell'intervallo per ogni punto del dominio è possibile ottenere un grafico che rappresenta le **bande di previsione** ottenute.

3.5 Verifica del modello mediante i residui

Anche le previsioni sono influenzate dalle ipotesi riguardo la distribuzione degli errori e dalla non variabilità della varianza. Può succedere che i dati analizzati non soddisfino queste ipotesi ed è necessario avere degli strumenti che permettono di identificare questi casi.

Il modo più semplice per verificare ciò è rappresentare con un grafico, ad esempio con un istogramma, la distribuzione dei residui per osservare se questa ha la classica forma campanulare della distribuzione normale.

Un altro grafico è il tracciamento dei residui rispetto la stima $\hat{y}$ ottenuta il quale aiuta a cogliere alcuni aspetti che ci aspettiamo vengano rappresentati dal modello. In particolare ci si aspetta di ottenere una nuvola di punti senza che ci siano dei pattern.

Prendere i grafici dalle slide nuove, sono fatti meglio

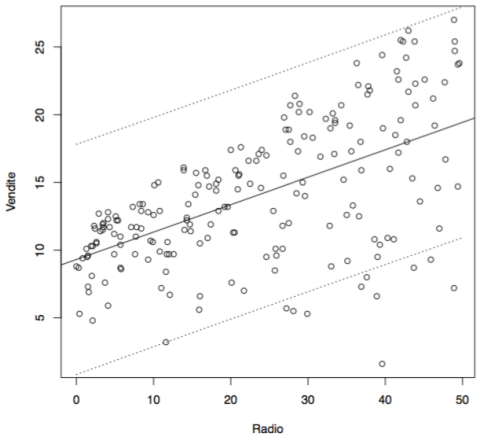


Figura 3.9: Bande di previsione

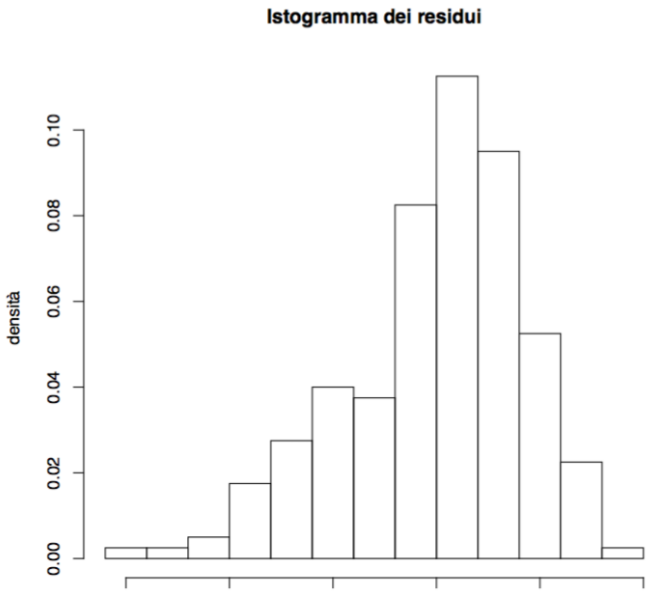


Figura 3.10: Istogramma dei residui per il data set delle vendite. In questo caso la campana non è proprio centrata, ma questo può essere dovuto a fattori casuali.

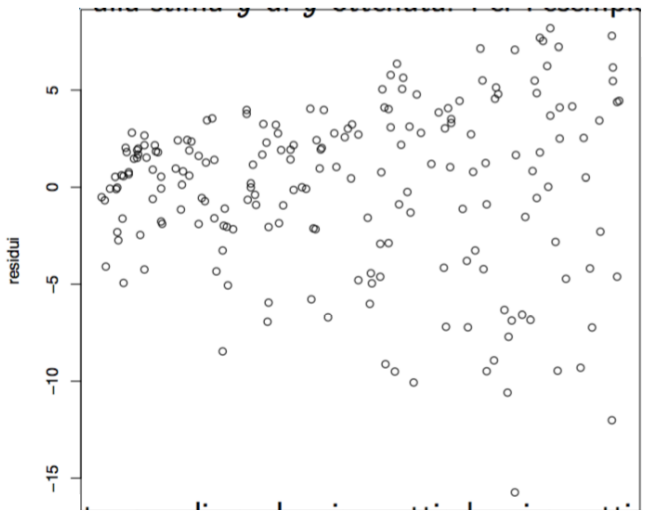


Figura 3.11: Residui rispetto $\hat{y}$, i punti sono sparsi quindi le assunzioni sembrano rispettate

In questo modo è possibile osservare se la varianza è costante o se ci sono degli effetti sistematici che il modello non ha colto, oppure se il modello utilizzato non è adeguato. Questi aspetti possono essere colti tracciando la dispersione sia per le x che per le y, il risultato che si ottiene è lo stesso, solo che viene sfasato di un fattore.

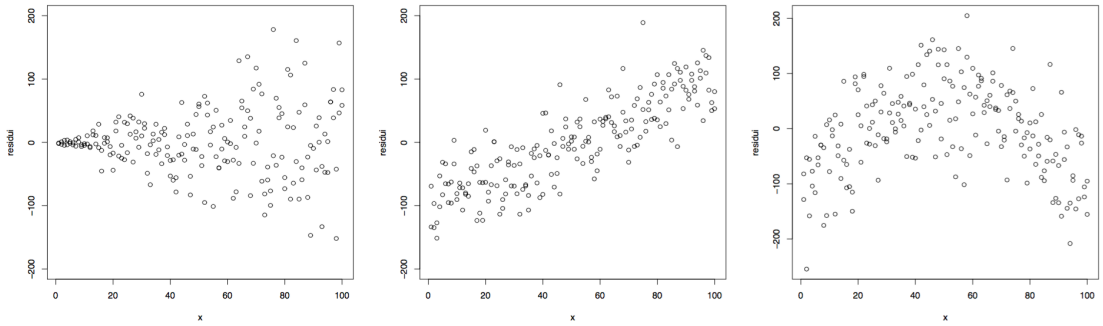


Figura 3.12: Da sinistra verso destra: Varianza non costante, componente sistematica non considerata dal modello, modello inadeguato. Tipicamente quando si verifica il caso centrale c'è un errore nei calcoli del modello.

Dai grafici è possibile notare anche la presenza di valori anomali (**outliers**), osservazioni errate che non rispettano la distribuzione, che devono essere trattate in modo particolare per evitare che influenzino in modo errato il modello.

Per verificare se la distribuzione è normale si può costruire un **qq-plot** dato che la forma dell'istogramma è influenzata dalla dimensione degli intervalli.

Come prima cosa viene creata la distribuzione dei residui per i campioni osservati in modo che ogni punto sia caratterizzato dalla percentuale di punti minori o uguali rispetto alla totalità dei dati. Questa percentuale viene stimata utilizzando le varie frequenze.

Dopodiché è possibile confrontare i quantili della distribuzione ottenuta con quelli della normale standardizzata.

Se per ciascuno punto i quantili sono circa uguali a quelli della normale, allora anche la distribuzione dei residui può essere considerata normale.

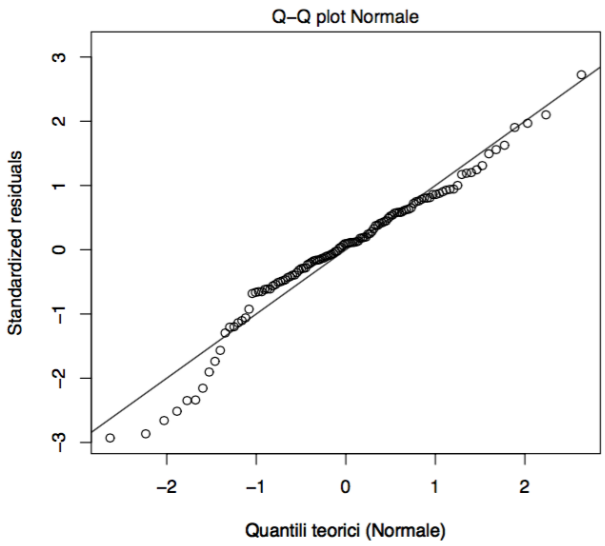


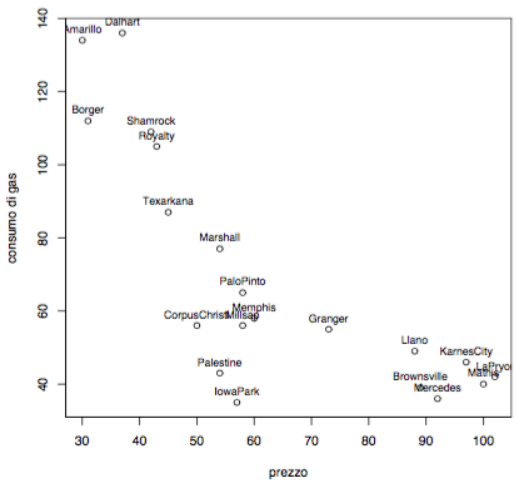
Figura 3.13: QQ-Plot, tanto più i punti sono vicini alla bisettrice, tanto più la distribuzione può essere considerata normale.

Capitolo 4

Il modello lineare nei parametri

Problema di riferimento: come il prezzo influenza il consumo di gas? Si hanno a disposizione le informazioni relative alla domanda di gas e al prezzo dello stesso per 20 città in Texas.

Si vuole riuscire a capire se c'è una correlazione tra le due cose.



4.1 Un primo modello lineare

Ipotizzando che ci sia una relazione lineare è possibile utilizzare il modello del capitolo precedente:

$$y = \alpha + \beta x + \epsilon$$

$$\hat{\beta} = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Utilizzando l'ambiente R si ottengono delle informazioni relative all modello ottenuto:

```
lm(formula = gas ~ prezzo)
Residuals:
    Min      1Q  Median      3Q     Max
```

```
-40.625 -10.719  -1.136  14.073  38.292
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  138.561     13.552  10.225 6.34e-09 ***
prezzo        -1.104      0.202  -5.467 3.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 20.86 on 18 degrees of freedom
Multiple R-Squared: 0.6241,     Adjusted R-squared: 0.6033
F-statistic: 29.89 on 1 and 18 DF,  p-value: 3.417e-05
```

Dai dati si può notare che l'indice R^2 è uguale a 0.62, il che indica un buon andamento lineare. Inoltre come *p-value* si ottiene un valore molto basso, il che porta a rifiutare l'ipotesi nulla.

Tracciando però i grafici dei residui è possibile osservare c'è una componente indipendente che non è lineare.

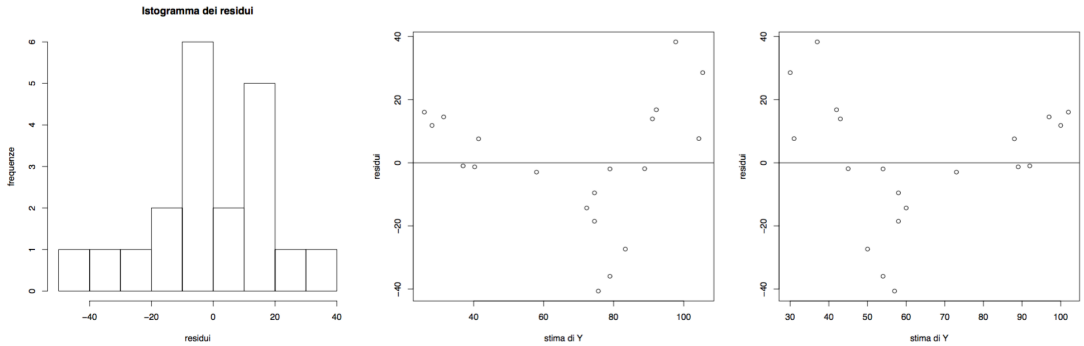


Figura 4.1: Tracciamento dei residui per il primo modello. È possibile notare la presenza di indipendente.

4.2 Considerazioni sul problema e un secondo modello

Considerando il problema modellato è possibile fare alcune osservazioni:

- il consumatore potrebbe destinare solamente un determinato budget κ per l'acquisto del gas, ovvero $x \cdot y = \kappa$.
- è ragionevole pensare che il consumatore debba consumare una quantità minima di gas γ
- essendo il mercato del gas regolamentato, c'è un prezzo minimo di 7 centesimi al metro cubo sotto il quale non è possibile vendere il gas.

Tenendo in considerazione quanto elencato si arriva ad avere l'equazione:

$$(x - 7)(y - \gamma) = \kappa$$

la quale può essere riscritta in un modo più simile a quella del modello lineare

$$y = \gamma + \kappa \cdot \frac{1}{x-7}$$

e sostituendo la variabile x con $z = \frac{1}{x-7}$, si ottiene proprio la stessa equazione la quale permette di calcolare la retta ai minimi quadrati.

Questo è possibile perché quello che finora è stato chiamato modello lineare è un caso particolare dei **modelli lineari nei parametri**. Ovvero la limitazione data dalla linearità non riguarda le variabili, ma riguarda solamente i **parametri** del modello.

Quando viene utilizzato il metodo dei minimi quadrati con questi modelli è necessario tenere in considerazione le trasformazioni che vengono fatte alle variabili, perché i valori calcolati ai minimi quadrati riguardano le variabili trasformate e non quelle di partenza, è necessario quindi **scalare** in modo opportuno i valori[1].

La formulazione più generale del modello lineare è quindi

$$g(y) = \alpha + \beta h(x) + \epsilon$$

Il modello ottenuto per la nuova formulazione è:

```
lm(formula = gas ~ I(1/(prezzo - 7)))
Residuals:
Min    1Q  Median  3Q    Max
-29.617 -4.574 2.394 7.800 30.917
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.918 8.376 0.468 0.646
I(1/(prezzo - 7)) 3034.938 357.037 8.500 1.02e-07 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 15.19 on 18 degrees of freedom
Multiple R-Squared: 0.8006, Adjusted R-squared: 0.7895
F-statistic: 72.26 on 1 and 18 DF, p-value: 1.022e-07
```

ovvero la retta

$$y = 3.918 + 3034.938 \cdot \frac{1}{x-7}$$

Dai dati del nuovo modello è possibile osservare che la varianza residua (`Residual standard error` elevato al quadrato) è passata da circa 391 a circa 207, ovvero il quadrato degli errori di previsione è stato ridotto di quasi il 50%. Lo stesso effetto può essere visto utilizzando R^2 che da 0.6241 passa a 0.8006.

[1] Se viene scalata solamente la x non c'è questo problema perché i minimi quadrati considerano solamente le distanze rispetto l'asse y.

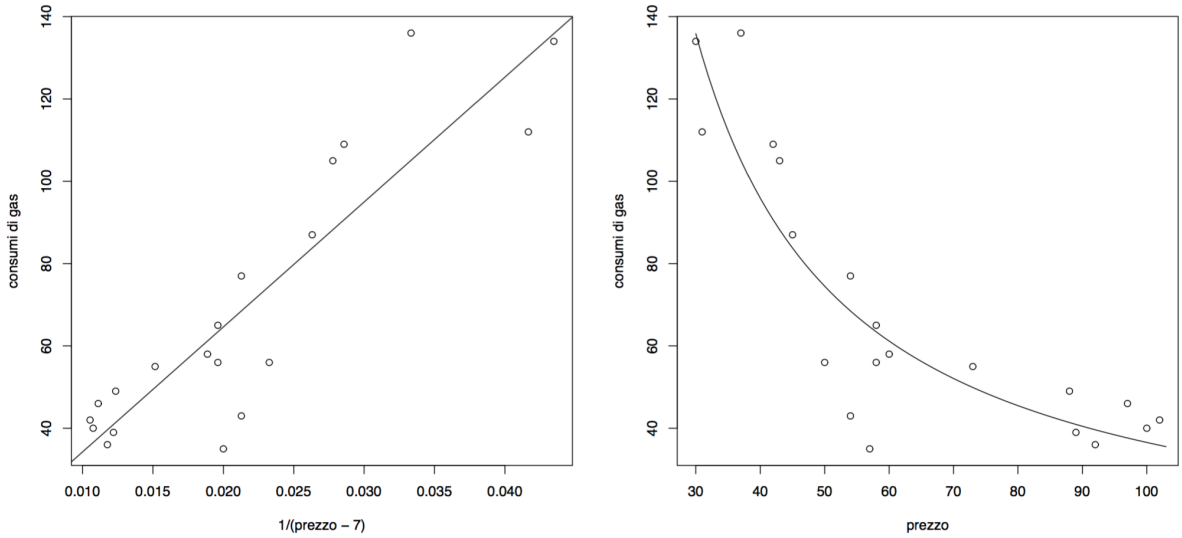


Figura 4.2: A destra la retta rispetto z. A sinistra il modello lineare tracciato rispetto x.

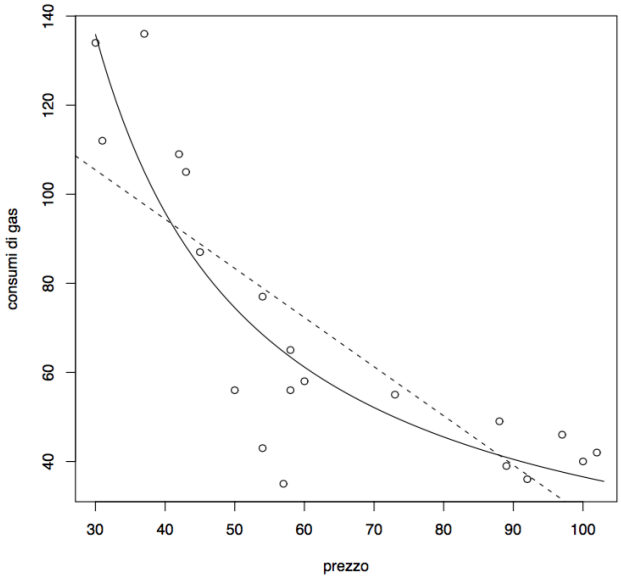


Figura 4.3: Confronto grafico tra i due modelli.

4.3 Modello lineare con trasformate

Cambia il dataset di riferimento, si vuole controllare se il reddito nazionale influisce sulla speranza di vita media dello stato.

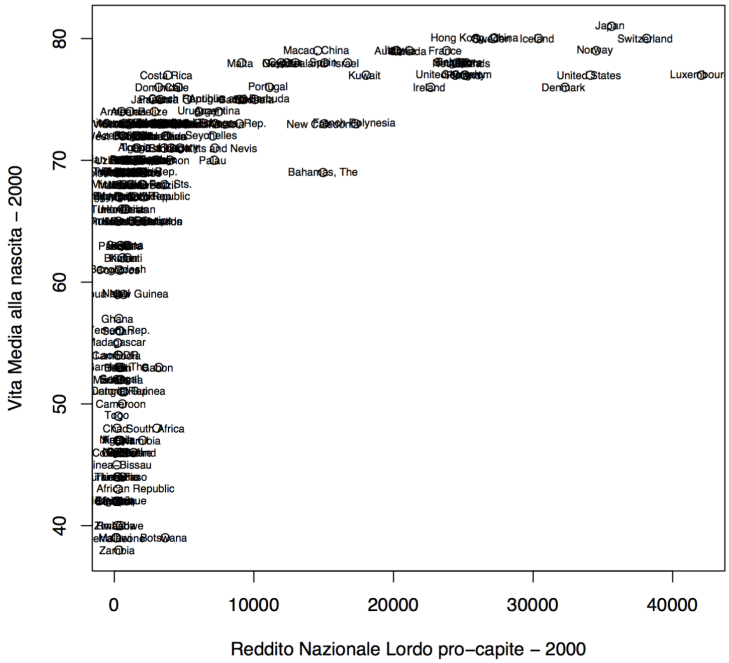


Figura 4.4: Dataset - GNI - ELF

La prima cosa da fare è osservare come si comporta il modello lineare senza trasformazioni:

```
lm(formula = elf ~ GNIpc)
Residuals:
Min    1Q  Median  3Q    Max
-24.924 -7.512 4.119 7.431 12.948
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.133e+01 8.967e-01 68.390 < 2e-16 ***
GNIpc 7.115e-04 8.230e-05 8.645 3.76e-15 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 9.903 on 171 degrees of freedom
Multiple R-Squared: 0.3041, Adjusted R-squared: 0.3
F-statistic: 74.73 on 1 and 171 DF, p-value: 3.757e-15
```

Si può notare come l'indice R^2 sia molto basso (0.3041), ma risulta essere molto significativo perché, per il *p-value* ottenuto sia ha che è improbabile che valga l'ipotesi nulla.

Tracciando il modello e il grafico dei residui è possibile notare che

- La retta ottenuta non curva abbastanza e quindi non si adatta bene ai dati

- Il modello prevede una vita media che può essere maggiore di 90 anni, il che è abbastanza improbabile.

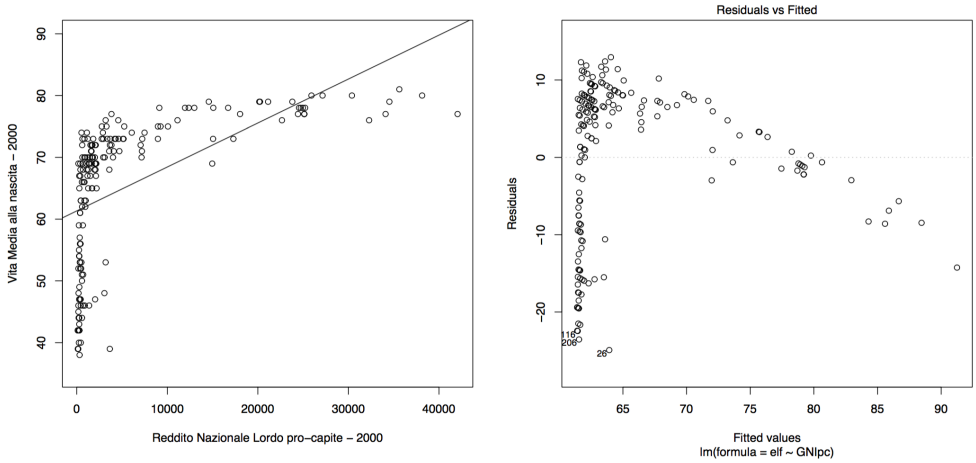


Figura 4.5: Primo modello e residui ottenuti

Per adattare meglio la curva è possibile utilizzare la scala logaritmica per l'asse delle x. In questo modo, al crescere del reddito viene dato via via meno peso. Inoltre, rappresentando graficamente questa trasformazioni si ottiene una nuvola di punti più simile ad una retta.

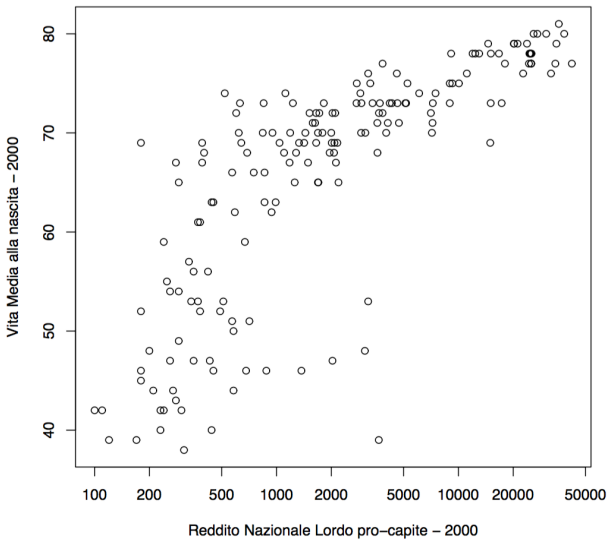


Figura 4.6: Grafico che utilizza la scala logaritmica per i valori delle x. L'asse è comunque etichettato con i valori originali.

Il modello diventa quindi

$$(\text{vita media alla nascita}) = \alpha + \beta \log(\text{reddito nazionale pro capite})$$

```
lm(formula = elf ~ I(logGDP), data = elf.data)
```

```
Residuals:
Min      1Q    Median    3Q      Max
-29.6591 -2.9511 0.7906 5.1050 17.4844
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.6701 2.8447 7.969 2.02e-13 ***
I(logGDP) 5.6767 0.3677 15.438 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.674 on 174 degrees of freedom
Multiple R-Squared: 0.578, Adjusted R-squared: 0.5756
F-statistic: 238.3 on 1 and 174 DF, p-value: < 2.2e-16
```

Con questo secondo modello si ottiene un indice R^2 doppio rispetto al precedente e questo può essere osservato anche nella rappresentazione grafica del nuovo modello.

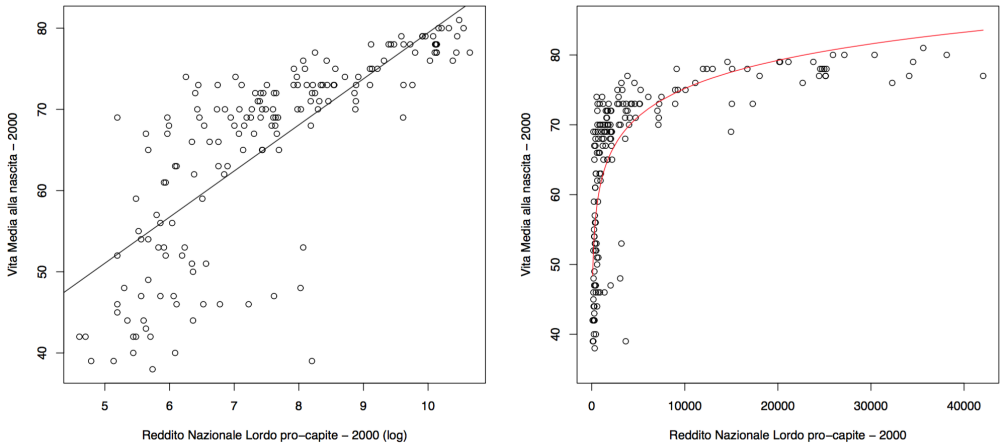


Figura 4.7: Secondo modello: a sinistra con la scala logaritmica, a destra normale.

C'è però ancora un problema che riguarda i valori estremi che non vengono approssimati bene dalla curva e lo si può notare anche dai residui.

Un'ulteriore modifica può essere quella di trasformare anche la variabile risposta, elevandola alla quinta, in modo da dare maggior peso ai valori maggiori.

Il modello ottenuto è quindi dato da

$$(\text{vita media alla nascita})^5 = \alpha + \beta \log(\text{reddito nazionale pro capite}) + \epsilon$$

Da notare che con questa formulazione le ipotesi sugli errori (media nulla, varianza costante, distribuzione normale, ecc.) **devono valere per gli errori su scala trasformata**.

Una volta calcolato il modello si ottiene

```
lm(formula = elf.5 ~ logGNI, data = elf.data)
Residuals:
Min          1Q          Median     3Q         Max
-1.801e+09 -2.817e+08 7.965e+06 3.036e+08 1.334e+09
Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

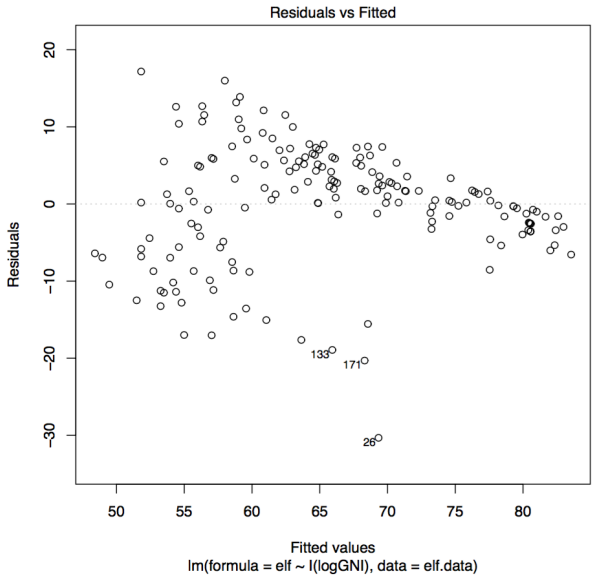


Figura 4.8: Residui per il secondo modello

```
(Intercept) -2.345e+09 1.832e+08 -12.80 <2e-16 ***
logGNI 5.164e+08 2.376e+07 21.73 <2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4.89e+08 on 171 degrees of freedom
Multiple R-Squared: 0.7341, Adjusted R-squared: 0.7326
F-statistic: 472.2 on 1 and 171 DF, p-value: < 2.2e-16
```

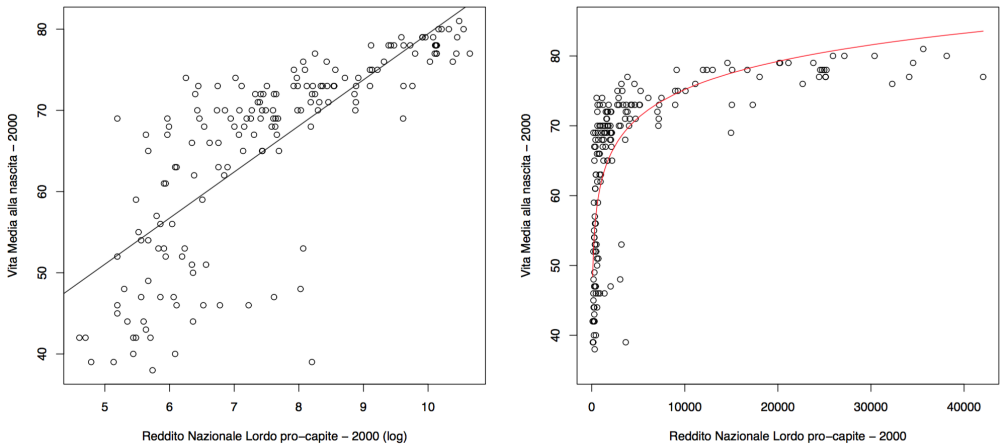


Figura 4.9: Terzo modello: a sinistra con la scala originale e a destra i residui.

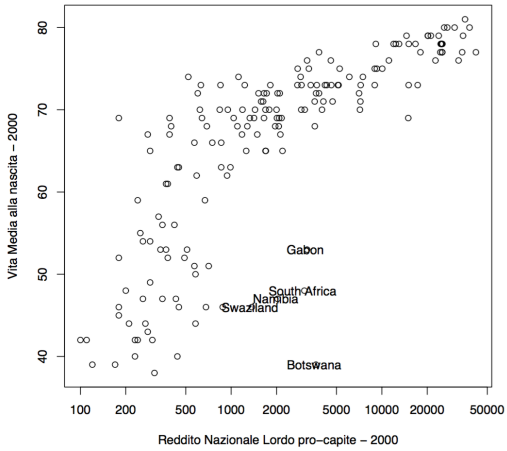
L'indice R^2 ottenuto fa riferimento ai residui calcolati sulla scala trasformata, quindi per avere un'indice di adattamento dei dati è possibile utilizzare la media dei quadrati dei residui per la variabile originale:

$$\frac{1}{n}\sum_{i=1}^{n}\left[(\text{vita media alla nascita})_i - \sqrt[5]{(\alpha + \beta\log(\text{reddito nazionale pro capite})_i)}\right]^2$$

Calcolando questo valore si ottiene 55.08, mentre con il secondo modello si aveva la varianza dei residui pari a 58.89. Si ottiene quindi una riduzione del 6% del quadrato degli errori di previsione.

4.4 Gestione degli outliers

I risultati ottenuti con il terzo modello sono molto buoni, ma osservando la distribuzione dei campioni si possono notare dei punti che escono dalla distribuzione.



Questi stati che si discostano sono principalmente africani e si può pensare che il comportamento diverso sia causato dalle loro condizioni socio-economiche, come l'AIDS che abbassa l'aspettativa di vita o il petrolio che aumentare la ricchezza solamente di poche persone.

Si può quindi pensare di calcolare il modello senza utilizzare questi outliers.

```
lm(formula = elf.5 ~ logGNI, data = elf.out1.data)
Residuals:
Min        1Q       Median      3Q        Max
-1.111e+09 -2.880e+08 -3.471e+07 2.817e+08 1.298e+09
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.344e+09 1.587e+08 -14.77 <2e-16 ***
logGNI 5.221e+08 2.059e+07 25.36 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 423200000 on 166 degrees of freedom
Multiple R-Squared: 0.7948, Adjusted R-squared: 0.7935
F-statistic: 642.9 on 1 and 166 DF, p-value: < 2.2e-16
```

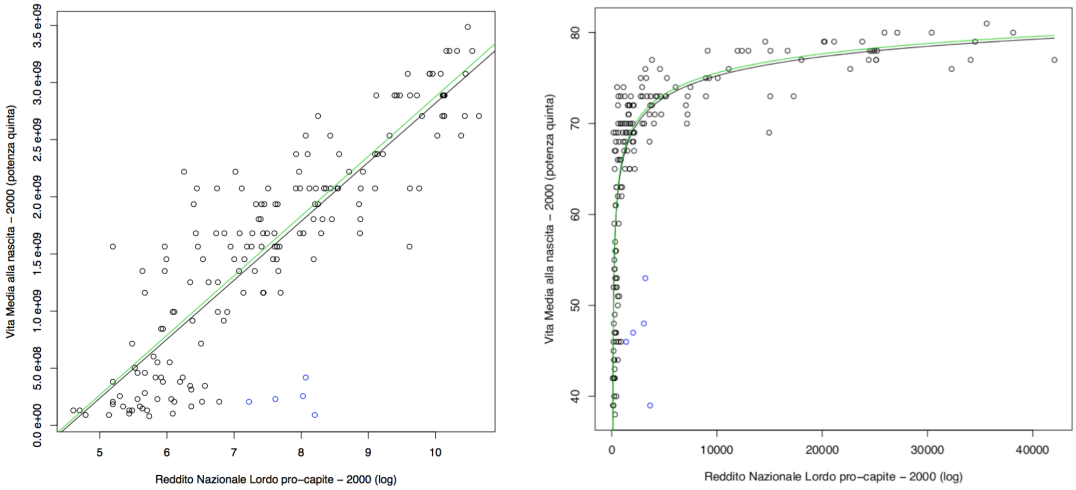


Figura 4.10: In verde il nuovo modello senza outliers, in nero il terzo modello. Si può notare che il nuovo modello è traslato verso l'alto perché non considera gli outliers (punti in blu)

In modo simile a come effettuati prima si può calcolare la varianza residua su scala originale, considera **solo** le osservazioni usate per la stima, ottenendo così una riduzione del 26%.

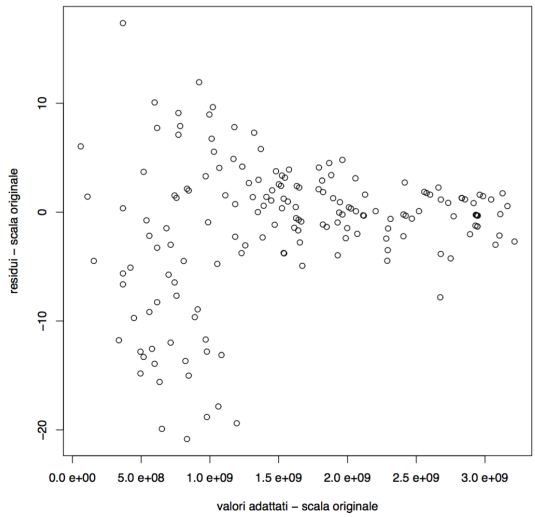


Figura 4.11: In residui nella scala originale. Quando l'ipotesi di indipendenza non è del tutto soddisfatta conviene effettuare una trasformazione della variabile risposta.

Capitolo 5

Modello lineare con più variabili esplicative

Dataset delle scuole: è ragionevole pensare che la quantità di soldi disponibili per una scuola sia una misura della qualità dell'insegnamento?

Variabili a disposizione:

- Spese medie annuali per studente
- rapporto studenti/insegnanti
- salario medio annuale
- percentuale di studenti che hanno effettuato il test SAT
- punteggio medio totale (per stato)
- punteggio per le materie verbali)
- punteggio per le materie matematiche

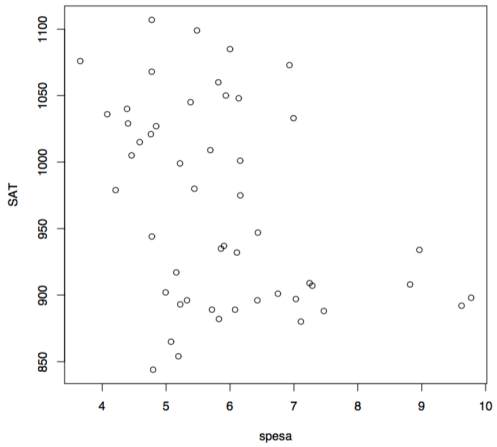


Figura 5.1: Diagramma di dispersione *spesa - SAT*

Per la distribuzione dei dati è possibile utilizzare la correlazione di **Bravais-Pearson**

$$\rho = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

il quale risulta essere -0.381 con *p-value* 0.006, quindi sembra che ci sia una correlazione lineare negativa tra le due variabili.

Per verificare ciò è possibile utilizzare il modello di regressione semplice

$$\text{SAT} = \alpha + \beta \text{SPESA} + \epsilon$$

Ottenendo $\hat{\alpha} = 1089.294$ e $\hat{\beta} = -20.892$ con *p-value* = 0.0064, quindi con i dati estremamente contrari all'ipotesi nulla.

p-value = probabilità di osservare un valore maggiore di t_{oss} rispetto la t di Student. In questo caso $t_{oss} = -2.85$.

Da questo primo modello emerge quindi che c'è una correlazione lineare tra queste due variabili, ma che l'aumento della spesa per l'istruzione porta ad una diminuzione della valutazione SAT, il che sembra contro intuitivo.

5.1 Correlazione parziale

Tra i dati a disposizione c'è anche la percentuale di studenti che si sono sottoposti al SAT e come si può notare dal grafico questa percentuale varia molto di stato in stato. Questo perché dall'esito del SAT dipende l'ammissione nelle università statali più prestigiose, mentre per altre università è possibile fare altri test. Morale della favola: negli stati con una percentuale di test effettuati minore l'esito medio è più alto perché lo fanno solo i secchioni.

Sarebbe bello poter definire un modello che tenga conto anche di questa correlazione.

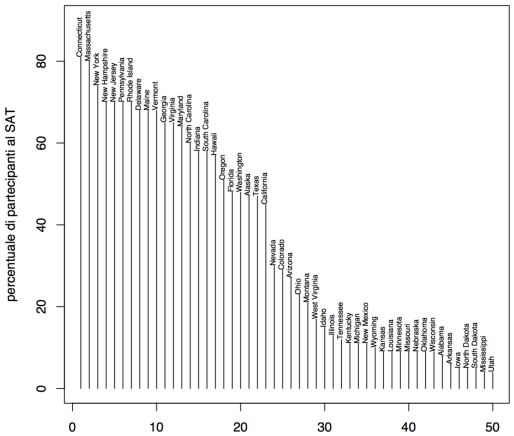


Figura 5.2: Percentuale di studenti che si sottopongo al SAT per stato

Una formulazione più generale di questo problema riguarda una variabile di risposta Y e almeno una variabile esplicativa X_1. È poi presente almeno un'altra variabile X_2 che potenzialmente influenza sia X_1 che Y.

Se c'è questa influenza, il coefficiente di correlazione ρ fornisce delle indicazioni errate in quanto X_2 ha un'effetto (assunto lineare) sulle altre variabili.

Il coefficiente di correlazione parziale

Si vuole stimare l'effetto lineare di X_2 sulle variabili di interesse.

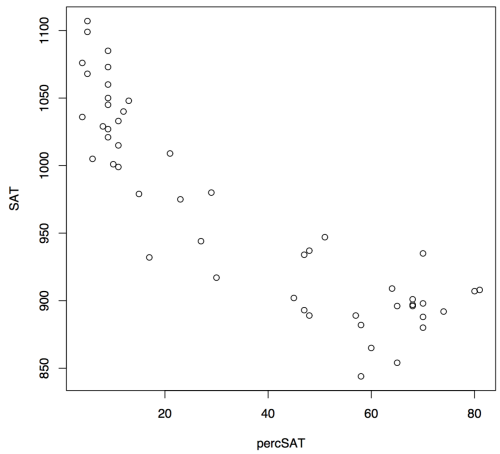


Figura 5.3: Diagramma di dispersione *perSAT - SAT*

Come prima cosa vengono calcolati i modelli

$$Y = \alpha_0 + \alpha_1 X_2 + \epsilon_Y$$
$$X_1 = \beta_0 + \beta_1 X_2 + \epsilon_{X_1}$$

con le solite ipotesi riguardo le distribuzioni degli errori.

Una volta ottenuti i modelli è possibile calcolare i residui, i quali misurano quanto della variabile Y e X_1 rispettivamente non è spiegato dalla variabile X_2.

$$Z_Y = Y - \alpha_0 - \alpha_1 X_2$$
$$Z_{X_1} = X_1 - \beta_0 - \beta_1 X_2$$

Con queste due nuove variabili è possibile calcolare il **coefficiente di correlazione parziale** tra Y e X_1 al netto dell'effetto lineare di X_2.

$$\rho_{YX_1:X_2} = \frac{cov(Z_Y, Z_{X_1})}{\sqrt{var(Z_Y)}\sqrt{var(Z_{X_1})}}$$

Calcolando questo coefficiente per il data set di riferimento si ottiene 0.391, ovvero il segno della correlazione tra *SAT* e *SPESA* è cambiato.

Stima del nuovo modello

$$\text{SAT} = \beta_0 + \beta_1 \text{SPESA} + \beta_2 \text{PERC_SAT} + \epsilon$$

Questo modello dovrebbe cogliere l'effetto congiunto delle due variabili, fornendo una misura di quanto ciascuna delle due variabili influisce singolarmente su *SAT*.

La stima viene fatta ai minimi quadrati, utilizzando la versione più generale del modello

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Si tratta quindi di stimare i coefficienti $\hat{beta}_0$, $\hat{beta}_1$ e $\hat{beta}_2$ in modo tale che

$$
\begin{aligned}
y_1 &\approx \hat{\beta}_0 + \hat{\beta}_1 x_{1,1} + \hat{\beta}_2 x_{2,1} + \epsilon \\
y_2 &\approx \hat{\beta}_0 + \hat{\beta}_1 x_{1,2} + \hat{\beta}_2 x_{2,2} + \epsilon \\
\vdots y_n &\approx \hat{\beta}_0 + \hat{\beta}_1 x_{1,n} + \hat{\beta}_2 x_{2,n} + \epsilon
\end{aligned}
$$

$$
s^2(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i})
$$

scegliendo $\hat{beta}_0$, $\hat{beta}_1$ e $\hat{beta}_2$ in modo da minimizzare s^2.

$$
s^2(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) \leq s^2(\beta_0, \beta_1, \beta_2)
$$

e con un po' di *math-magic* si ottiene:

$$
\begin{cases}
\hat{\beta}_1 & = \frac{cov(X_1,Y)sqm(X_2) - cov(X_2,Y)cov(X_1,X_2)}{sqm(X_1)sqm(X_2) - cov^2(X_1,X_2)} \\
\hat{\beta}_2 & = \frac{cov(X_2,Y)sqm(X_1) - cov(X_1,Y)cov(X_1,X_2)}{sqm(X_1)sqm(X_2) - cov^2(X_1,X_2)} \\
\hat{\beta}_0 & = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2
\end{cases}
$$

5.2 Notazione matriciale

Utilizzando l'algebra matriciale si riesce a semplificare di molto la formulazione precedentemente presentata.

Assumendo di avere un modello lineare con:

- $p - 1$ variabili esplicative
- un vettore $\vec{y}$ di dimensione n contenente le osservazioni della variabile risposta
- un vettore $\vec{\epsilon}$ n contenente le determinazioni della variabile aleatoria degli errori
- un vettore $\vec{\beta}$ di dimensione p dei parametri da stimare (un parametro per ogni variabile più uno per l'intercetta)
- una matrice X di dimensione $n \times p$ con le osservazioni delle variabili esplicative a partire dalla seconda colonna. La prima colonna è contiene tutti 1. Ogni riga rappresenta un'osservazione, ogni colonna rappresenta una singola variabile esplicativa.

$$
\mathbf{X} = \begin{bmatrix}
1 & x_{1,1} & x_{2,1} & \cdots & x_{p-1,1} \\
1 & x_{1,2} & x_{2,2} & \cdots & x_{p-1,2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{1,n} & x_{2,n} & \cdots & x_{p-1,n}
\end{bmatrix}
$$

Utilizzando questa notazione è possibile scrivere il modello lineare come

$$
\vec{y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}
$$

dove le variabili aleatorie del vettore $\vec{\epsilon}$ hanno media 0 e varianza σ^2 intercorrelate tra loro, ovvero la matrice delle variane e covarianze di $\vec{\epsilon}$ è diagonale con tutti gli elementi sulla diagonale uguali a σ^2.

Il problema dei minimi quadrati risulta quindi definito come

$$s^2(\vec{\beta}) = (\vec{y} - \mathbf{X}\vec{\beta})^T(\vec{y} - \mathbf{X}\vec{\beta})$$

e la soluzione viene trovata risolvendo l'equazione

$$\mathbf{X}^T\mathbf{X}\vec{\beta} = \mathbf{X}^T\vec{y}$$

ovvero

$$\vec{\hat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y}$$

La stima dei valori viene fatta quindi con

$$\vec{\hat{y}} = \mathbf{X}\vec{\hat{\beta}}$$

da notare che $\vec{\hat{y}}$ ha dimensione $n \times 1$, dove n è il numero delle righe della matrice $\boldsymbol{X}$, quindi se la matrice ha una sola riga, si ottiene un valore scalare.

5.3 Verifica del modello

Come prima cosa è necessario calcolare la varianza degli stimatori

$$\mathrm{var}(\vec{\hat{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

dove σ^2 è la varianza scalare di ciascuna ϵ_i. La varianza però non è nota, quindi è necessario effettuare una stima.

Capire se è un vettore o no

$$\widehat{\mathrm{var}}(\vec{\hat{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}s^2$$

con

$$s^2 = \frac{1}{n-p}(\vec{y}^T\vec{y} - \vec{\hat{\beta}}^T\mathbf{X}^T\vec{y})$$

5.3.1 Scomposizione della varianza

Nel caso della regressione semplice la *varianza totale* di y può essere scomposta nella somma della *varianza spiegata dal modello* e della *varianza residua*. Anche in questo caso è possibile fare lo stesso ragionamento.

Per semplicità viene considerato solo il numeratore della varianza, spesso nominato come **devianza** o *somma dei quadrati degli scarti della media*. In questo modo si possono tralasciare n e i gradi di libertà.

Si ha quindi che

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{devianza}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{devianza spiegata dal modello}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{devianza residua}}$$

Dando un po' di nomi ai pezzi si ottiene:

- $\sum_{i=1}^{n}(y_i - \bar{y})^2$: somma dei quadrati degli scarti della media o **devianza totale** (**SST**)
- $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$: somma dei quadrati degli scarti dalla regressione o **devianza spiegata dal modello** (**SSR**)
- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: somma dei quadrati degli scarti residui o **devianza residua** (**SSE**)

In modo analogo, anche i gradi di libertà sono scomponibili nella somma di due valori

$$\text{gdl totali} = \text{gdl della regressione} + \text{gdl dei residui}$$

che nel caso della regressione multipla con p variabili esplicative sono:

$$(n) = (p) + (n - p)$$

5.3.2 Test di significatività globale

Una prima verifica d'ipotesi che si può fare è quella sull'ipotesi che il modello lineare non serva per spiegare la relazione con l'alternativa, la quale dice che almeno a qualcosa serve.

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\ H_1 : \text{almeno un } \beta_j \neq 0 \quad j = 1 \ldots p \end{cases}$$

la verifica dell'ipotesi si fa al solito modo, utilizzando F_{oss} che dipende da R^2:

$$\begin{aligned} F_{oss} &= \frac{R^2}{1 - R^2} \frac{n - p}{p} \\ &= \frac{\text{devianza spiegata}}{\text{devianza residua}} \cdot \frac{n - p}{p - 1} \\ &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \cdot \frac{n - p}{p - 1} \\ &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 / (p - 1)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / (n - p)} \end{aligned}$$

Riassumendo

Con la notazione MSR ci si riferisce alla somma dei quadrati degli scarti delle regressione (SSR) divisi i gradi di libertà della regressione, ottenendo quindi:

$$MRS = \frac{SSR}{p}$$

In modo analogo

$$MST = \frac{SST}{n} \quad \text{e} \quad MSE = \frac{SSE}{n - p}$$

ovvero MST coincide con la varianza totale.

5.4 Gestione delle variabili esplicative

Dataset di riferimento informazione sui distretti di Boston, si vuole far previsione sul costo medio della casa.

Tracciando i vari diagrammi di dispersione per le variabili esplicative del dataset è possibile notare se queste sono tra loro correlate.

Una volta notata una sorta di correlazione è possibile provare ad utilizzare tutte le variabili per calcolare il modello lineare:

$$(\text{valore}) = \beta_0 + \beta_1(\text{crimine}) + \ldots + \beta_1 3(\text{poveri}) + \text{errore}$$

C'è però da notare una cosa, il dataset contiene delle variabili qualitative, come *fiume* che può prendere come valore *0 o 1* e *str* che ha 9 possibili valori.

Per poter utilizzare il modello lineare è necessario trasformare queste variabili discrete che possono assumere k valori in $k-1$ variabili **indicatrici** o **dummies**, una per ogni possibile valore (meno uno), che possono assumere come valore *0 o 1* e tali che al massimo una sola di queste variabili sia a 1.

Viene utilizzata una variabile in meno rispetto al numero dei valori perché il valore escluso (che viene fissato a priori) viene rappresentato quando tutte le altre variabili sono a 0.

La scelta del valore influisce sul significato dei parametri. Ovvero fissato un parametro di default (con tutte le variabili indicatrici a 0) i valore degli altri parametri esprire la variazione rispetto al parametro di default.

Ad esempio considerando reddito regione. Se la regione di default è il Veneto, gli altri parametri associati alle variabili dummies indicando la variazione del reddito in base a quello del veneto: le regioni più povere avranno un parametro negativo, mentre quelle più ricche del veneto avranno un parametro positivo.

Tornando ad dataset delle case, si ottiene quindi una matrice $\mathbf{X}$ con 506 righe e 21 colonne.

5.4.1 Verifica d'ipotesi per i coefficienti

In modo simile a come avviene con il modello a due parametri è possibile formulare le ipotesi:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

che vengono verificate utilizzando

$$\begin{aligned} t_{oss}(\beta_i) &= \frac{\hat{\beta}_i}{\sqrt{\dfrac{s^2}{\sum_{j=1}^n (x_{ij} - \bar{x}_i)^2}}} = \frac{\hat{\beta}_i}{\sqrt{\dfrac{\sum_{j=1}^n (y_{ij} - \hat{y}_j)^2/(n-p)}{\sum_{j=1}^n (s_{ij} - \bar{x}_i)^2}}} \\ &= \frac{\vec{a}^T \hat{\beta}}{s\sqrt{\vec{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \vec{a}}} \end{aligned}$$

dove $\vec{a}$ è un vettore con tanti elementi quanti sono i parametri, che contiene tutti 0 tranne nell'elemento i-esimo.

La statistica t_{oss} sotto H_0 si distribuisce come una t di Student con $n-p$ gradi di liberà. Nel caso d'esempio si ha $n-p = 506-21 = 485$. Da notare che p tiene conto anche di β_0

Da verificar che $n - \#var - 1 = n - p$.

5.4.2 Adattamento complessivo del modello

Come prima si può calcolare R^2 come

$$R^2 = \frac{\hat{\beta}^T \mathbf{X}^T \vec{y} - n\bar{y}^2}{\vec{y}^T \vec{y} - n\bar{y}^2}$$

La significatività di R^2 può essere valutata attraverso

$$\begin{aligned} F_{oss} &= \frac{\text{media dei quadrati degli scarti spiegati}}{\text{media dei quadrati degli scarti residui}} \\ &= \frac{\hat{\beta}^T \mathbf{X}^T \vec{y} - n\bar{y}^2/(p-1)}{\vec{y}^T \vec{y} - n\bar{y}^2/(n-p)} \end{aligned}$$

Sotto H_0 F_{oss} si distribuisce come una F di Snedecor con $p-1$ gradi di libertà al numeratore e $n-p$ gradi di libertà al denominatore

Nelle slide c'è n-1 al posto di p-1, solo nella formula

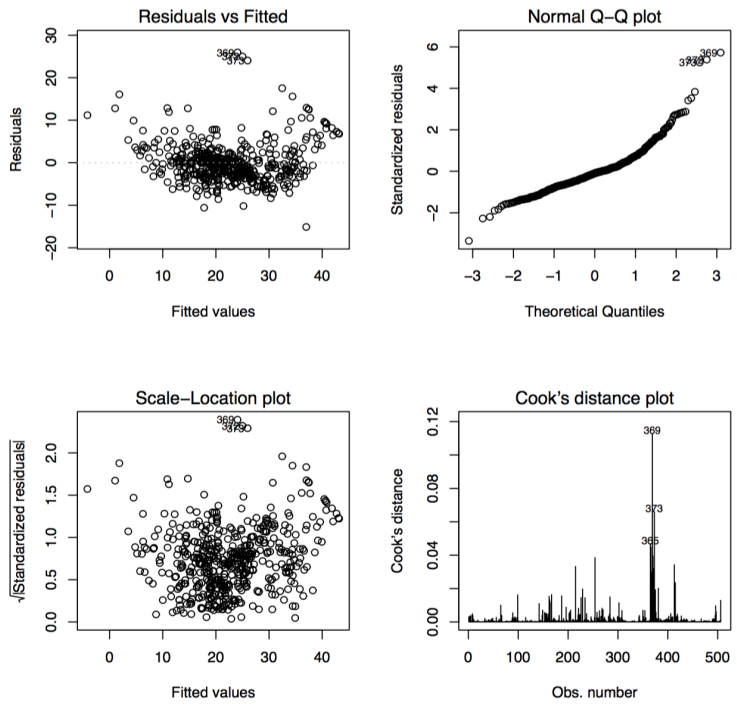
5.4.3 Valutazione del primo modello

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      35.259615   5.433641   6.489 2.14e-10 ***
crimine          -0.108821   0.032685  -3.329 0.000937 ***
large.lots        0.054896   0.014149   3.880 0.000119 ***
perc.nonaffitti   0.023760   0.063648   0.373 0.709081 <- variabile non rilevante per prevedere la
fiume1            2.524163   0.863181   2.924 0.003614 **        risposta
conc.NO         -17.573132   3.896362  -4.510 8.13e-06 ***
n.stanze          3.665491   0.421196   8.703  < 2e-16 ***
perc.prima80      0.000461   0.013227   0.035 0.972210
dist.lavoro      -1.554546   0.201909  -7.699 7.75e-14 ***
strade2           1.488905   1.477509   1.008 0.314095
strade3           4.681254   1.335296   3.506 0.000498 ***
strade4           2.576234   1.187369   2.170 0.030514 *
strade5           2.918493   1.207610   2.417 0.016028 *
strade6           1.185839   1.464013   0.810 0.418342 <- variabile comunque rilevante per via
strade7           4.878992   1.571198   3.105 0.002012 **  delle altre strade (al massimo si può
strade8           4.839836   1.491657   3.245 0.001257 **  dire che la strada6 è simile alla
strade24          7.461674   1.788679   4.172 3.58e-05 *** strada 1)
tasse            -0.008748   0.003895  -2.246 0.025144 *
stud.insegn      -0.972419   0.144478  -6.731 4.78e-11 ***
perc.neri         0.009394   0.002661   3.531 0.000454 ***
perc.poveri      -0.529226   0.050640 -10.451  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.694 on 485 degrees of freedom
Multiple R-Squared: 0.7499,     Adjusted R-squared: 0.7396
F-statistic:  72.7 on 20 and 485 DF,  p-value: < 2.2e-16
                      p-1     n-p
```

Come si può notare il modello ottenuto risulta essere abbastanza buono, tuttavia:

- Nel grafico dei residui si nota una forma ad U e la presenza di alcuni valori particolari
- Le code del QQ-Plot si allontanano dalla distribuzione normale



- Ci sono dei valori più significativi che influenzano il modello

Si può quindi provare ad applicare delle trasformazioni delle variabili e ad omettere i valori troppo influenti, in modo da ottenere un R^2 maggiore.

Da notare che in questo caso R^2 può essere confrontato con quello precedentemente ottenuto perché le trasformazioni non riguardano la variabile risposta.

5.4.4 Analisi della varianza per la scelta delle variabili esplicative

Avendo a disposizione tante variabili può essere che ce ne siano alcune di non significative per il modello.

Serve quindi un meccanismo formale per cercare e rimuovere queste variabili, tenendo conto che l'inclusione o l'esclusione di una singola variabile influisce sulla varianza dei residui. Questa procedura prende il nome di **analisi della varianza**.

Si è già visto come scomporre la varianza:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{devianza}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{devianza spiegata dal modello}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{devianza residua}}$$

Nel caso della regressione multipla è possibile ottenere la stessa cosa con

$$\underbrace{\vec{y}^T\vec{y}}_{\textbf{SST}\text{:somma dei quardati totale}} = \underbrace{(\vec{y}-\vec{\bar{y}})^T(\vec{y}-\vec{\bar{y}})}_{\textbf{SSE}\text{:somma dei quadrati degli errori}} + \underbrace{\vec{\hat{\beta}}\mathbf{X}^T\vec{y}}_{\textbf{SSR}\text{:somma dei quadrati della regressione}}$$

e F_{oss} per la validità globale del modello può essere espressa come

$$F_{oss} = \frac{SSR/p}{SSE/(n-p)}$$

Una volta scomposta la varianza è possibile calcolare la tabella **ANOVA**, ovvero **AN**alysis **O**f **VA**riance.

È possibile quindi costruire la tabella ANOVA in seguente

Fonte	SS	gdl	MS
β_1	$SSR(\beta_1)$	q_1	$SSR(\beta_1)/q_1$
$\beta_2\|\beta_1$	$SSR(\beta_2\|\beta_1)$	q_2	$SSR(\beta_2\|\beta_1)/q_2$
$\beta_3\|\beta_1,\beta_2$	$SSR(\beta_3\|\beta_1,\beta_2)$	q_3	$SSR(\beta_3\|\beta_1,\beta_2)/q_3$
...	...	...	...
Errore	$SSE = \mathbf{y}^T\mathbf{y} - \hat{\beta}\mathbf{X}^T\mathbf{y}$	$n-p$	$SSE/(n-p)$
Totale	$SST = \mathbf{y}^T\mathbf{y}$	n	

- Applicando questa procedura per verificare ogni singolo β_i, dopo l'inclusione dei termini adattati precedentemente, si ottiene la statistica test

$$F_{oss} = \frac{MSR(\beta_i|\text{precedenti componenti di } \beta)}{MSE}$$

che sotto H_0 si distribuisce come una $F_{1,n-p}$ (data l'inclusione nel modello delle precedenti componenti β_j).

- Si consideri che l'ordine con cui vengono aggiunti termini nel modello diventa ora importante. Cioè, ad esempio, se β è partizionato in β_1, β_2 e β_3, $SSR(\beta_3|\beta_1)$ sarà in genere diverso da $SSR(\beta_3|\beta_2)$, e il testo per l'inclusione di β_3 potrebbe dare risultati diversi (una volta accettare e una volta rifiutare) a seconda che β_1 o β_2 sia stimato prima.

- La tabella ANOVA per il nostro dataset sui prezzi delle case dipende, come abbiamo visto dall'ordine con cui inseriamo i dati:
- un primo ordine potrebbe essere: crimine, large.lots, indus, fiume, perc.prima40, strade, tasse, stud.insegn, perc.neri, conc.NO^2, n.stanze^2, log(dist.lavoro), log(perc.poveri)
- la tabella di analisi della varianza (a cui abbiamo aggiunto la statistica test F_{oss} e il p-value) risulta

```
                  gdl   SS       MS       F        Pr(>F)
crimine            1  6856.0  6856.0 637.1910 < 2.2e-16 ***
large.lots         1  3919.9  3919.9 364.3163 < 2.2e-16 ***
indus              1  3257.0  3257.0 302.7015 < 2.2e-16 ***
fiume              1   646.6   646.6  60.0937 5.525e-14 ***
perc.prima40       1    80.7    80.7   7.4963  0.006414 **
strade             8  2754.4   344.3  31.9990 < 2.2e-16 ***
tasse              1   429.1   429.1  39.8796 6.202e-10 ***
stud.insegn        1  2596.6  2596.6 241.3297 < 2.2e-16 ***
perc.neri          1   453.2   453.2  42.1208 2.158e-10 ***
conc.NO^2          1   532.5   532.5  49.4856 6.992e-12 ***
n.stanze^2         1 10447.6 10447.6 970.9934 < 2.2e-16 ***
log(dist.lavoro)   1   616.9   616.9  57.3347 1.925e-13 ***
log(perc.poveri)   1  1147.1  1147.1 106.6071 < 2.2e-16 ***
Errori           476  5121.6    10.8
---------------------------------------------------------
Totale           496 38859.2
---
Codici di Sign.: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendice A

Laboratorio

A.1 Un po' di cose su R

- Tutti gli oggetti sono vettori
- `ls()` per vedere le variabili disponibili
- `x <- c(2,3,4,5)` crea un vettore con 1,2,3,4.
- notazione `[1:20]` per un vettore con la successione da 1 a 20
- `xx <- seq(from=100, to=1)` crea sempre una sequenza di numeri, con parametro opzionale `by` per specificare lo step
- `rep(2,5)` crea un vettore con 5 elementi uguali a 2
- `a <- c(rep(2,3),4,5,rep(1,5))`, `a = 2 2 2 4 5 1 1 1 1 1`
- `2*x` esegue il prodotto scalare
- `length(x)` per la lunghezza del vettore
- `max(x)` e `min(x)`
- `sum(x)` che ritorna un vettore di un solo elemento con la somma
- `mean(x)`, `var(x)`, `range(x)`
- `x[7]` per estrarre il settimo elemento di `x`, l'indice credo parta da 1
- `x[-4]` ritorna un vettore senza il quarto elemento
- `x <- matrix(c(2,3,5,7,11,13),nrow = 3)` crea una matrice con gli elementi specificati e 3 righe. Alternativamente è possibile specificare anche il numero di colonne.
- `x2 <- scan("nome file", sep="")` con `sep` opzionale, per caricare il contenuto di un file in un vettore, per caricare una matrice `x2 <- matrix(scan(...), ncol = 3, byrow=TRUE`.
- `str(x)` specifica la struttura dell'oggetto
- `dim(x)` ritorna la dimensione di una matrice, se invocato con un vettore ritorna `NULL`.
- `x[18,]` per ottenere la 18-esima riga di una matrice
- **Dataframe**: matrice le cui colonne possono avere formati diversi

- `ciliegi <- read.table("nome file")`.
- `names(ciliegi)` è il vettore con i nomi delle colonne del dataframe
- `names(ciliegi) <- c("diametro", "altezza", "volume")` permette di impostare il nome delle colonne, può anche essere specificato come parametro opzionale `col.names` della funzione `read.table`.
- `summary(ciliegi)` fornisce degli indicatori per ciascuna colonna
- **Mediana**: elemento centrale di una distribuzione ordinata in senso crescente, **primo e terzo quartile**: generalizzazione della mediana, rispettivamente l'elemento che sta al 25 e 75 per cento della distribuzione. La differenza tra i due quartili da l'idea di quanto è variabile la distribuzione.
- I dataframe possono essere acceduti anche con il nome della colonna `ciliegi$volume`.
- **attach di un file**: aggiungere al workspace un oggetto, ovvero `attach(ciliegi)` permette di accedere al nome della colonna direttamente utilizzando `volume`. Come complementare c'è il comando `detach`.
- `hist(diametro)` crea l'istogramma per il diametro
- `help(hist)` per avere l'help di una funzione
- l'istrogramma che viene generato di default può contenere dei buchi, conviene quindi adattare il numero di colonne utilizzando il parametro `breaks`
- `boxplot(diametro)` fornisce il box plot di un valore, è un grafico che rappresenta la mediana, i quartili e il 5 e 95%. Risulta più espressivo dell'istogramma. L'ampiezza della scatola rappresenta la variabilità dei dati.
- `ciliegi[altezza>80,]` prende tutti i ciliegi con altezza maggiore di 80.
- `library(MASS)` permette di caricare la libreria MASS
- `search()` permette di visualizzare la lista degli ottetti in cui R va a cercare quando deve eseguire un comando
- Gli attributi qualitativi vengono trattati come tipo Factor
- `table(painters$School)` crea la tabella con le frequenze delle varie qualità
- `barplot(..)` fa il plot delle barre per una variabile discreta
- `pie(...)` fa il grafico a torta, anche se è sconsigliabile utilizzare un grafico a torta perché per l'occhio umano fa fatica a vedere la differenza tra gli angoli.
- come scale colori si possono utilizzare `heat.colors(k)`, `rainbow(k)`, ...
- `plot(x,y)` disegna un diagramma di dispersione, il parametro `pch` specifica il tipo di carattere, `pch=16` rappresenta i pallini pieni, `col` specifica il colore da utilizzare, possono indicare `col=painter$School` per far variare il colore in base al valore dell'attributo quantitativo

Appendice B

Formulario

B.1 Formule generiche

$$\begin{aligned}\mathrm{var}(X) &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i^2+\bar{x}^2-2x_i\bar{x})}{n} = \frac{\sum_{i=1}^{n}x_i^2+n\bar{x}^2-2\bar{x}\sum_{i=1}^{n}x_i}{n}\\ &= \frac{\sum_{i=1}^{n}x_i^2}{n}-\bar{x}^2\end{aligned}$$

$$\begin{aligned}\mathrm{cov}(X,Y) &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n} = \frac{\sum_{i=1}^{n}(x_iy_i-x_i\bar{y}-\bar{x}y_i+\bar{x}\bar{y})}{n} = \frac{\sum_{i=1}^{n}x_iy_i-\bar{y}\sum_{i=1}^{n}x_i-\bar{x}\sum_{i=1}^{n}y_i+n\bar{x}\bar{y}}{n}\\ &= \frac{\sum_{i=1}^{n}(x_iy_i)}{n}-\bar{x}\bar{y}\end{aligned}$$

B.1.1 Coefficiente di correlazione di Pearson

$$\rho_{YX_1:X_2} = \frac{\mathrm{cov}(Z_Y, Z_{X_1})}{\sqrt{\mathrm{var}(Z_Y)\mathrm{var}(Z_{X_1})}}$$

$$-1 \leq \rho_{YX_1:X_2} \leq 1$$

Se $\rho_{YX_1:X_2} = 0$ le variabili Y e X_1 non sono correlate, se invece vale 1 o -1 sono correlate (direttamente o indirettamente).

B.1.2 Scomposizione della varianza

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SST: devianza}} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{SSE: devianza residua}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{SSR: devianza spiegata dal modello}}$$

$$\underbrace{\vec{y}^T\vec{y}}_{\textbf{SST}\text{:somma dei quardati totale}} = \underbrace{(\vec{y}-\vec{\bar{y}})^T(\vec{y}-\vec{\bar{y}})}_{\textbf{SSE}\text{:somma dei quadrati degli errori}} + \underbrace{\vec{\hat{\beta}}\mathbf{X}^T\vec{y}}_{\textbf{SSR}\text{:somma dei quadrati della regressione}}$$

B.2 Modello lineare

$$S^2(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\hat{\beta}_1 = \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)} = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

B.2.1 Residui

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$\operatorname{var}(R) = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \operatorname{var}(Y) - \frac{\operatorname{cov}^2(X,Y)}{\operatorname{var}(X)}$$

B.2.2 Coefficiente di determinazione

$$R^2 = 1 - \frac{\operatorname{var}(R)}{\operatorname{var}(Y)}$$

B.2.3 Stima della varianza della distribuzione

La varianza degli errori può essere approssimata con la varianza dei residui, scalata per un opportuno fattore, che potrebbero essere i gradi di libertà della distribuzione dei residui. Nel caso della regressione lineare a due parametri si ha che i gradi di libertà sono $n - p = n - 2$

$$\begin{aligned}
\hat{s}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(r_i - 0)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \frac{SSE}{n-p}
\end{aligned}$$

$$\operatorname{var}(\hat{\beta}_0) = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = s^2\left(\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

$$\operatorname{var}(\hat{\beta}_1) = s^2\left(\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Varianza della stima, ha gli stessi gradi di libertà della varianza dei residui $n - p$:

$$\widehat{\operatorname{var}(\hat{y}_0)} = s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

B.2.4 Verifica delle ipotesi

t_{oss} per $\hat{\beta}_1$

$$t_{oss} = \frac{(\hat{\beta}_1 - 0)}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

$$\left[\hat{\beta}_1 \pm t_{n-2,1-\alpha/2}\sqrt{\frac{s^2}{\sum_{i=i}^{n}(x_i - \bar{x})^2}}\right]$$

t_{oss} per $\hat{\beta}_0$

$$t_{oss} = (\hat{\beta}_0 - \beta') \Big/ s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\left[\hat{\beta}_0 \pm t_{n-2,1-\alpha/2} \cdot s \cdot \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

livello di significatività osservato o **p-value**: indicatore uguale alla probabilità di osservare, supponendo vera H_0, un valore di t_{oss} uguale o "più lontano" da H_0 di quanto effettivamente è osservato, ovvero costituisce una misura di quanto l'ipotesi nulla è plausibile sulla base dei dati osservati.

$$p - value = 2 \times P(t\text{con } n - p \text{ gradi} \geq t_{oss})$$

Per valutare l'andamento globale del modello è possibile utilizzare F_{oss} che si comporta in modo simile a t_{oss} con la differenza che la distribuzione di riferimento è la F di Snedecor con $p - 1$ gradi di libertà al numeratore e $n - p$ gradi di libertà al denominatore.

$$F_{oss} = \left(\frac{R^2}{1 - R^2}\right)\left(\frac{n - k}{k - 1}\right)$$

$$k = \text{numero di parametri del modello}$$

Regola di accettazione/rifiuto

1. Fisso un intervallo di confidenza $1 - \alpha$
2. Calcolo, utilizzando le tavole $t_{n-p,1-\alpha/2}$
3. Calcolo il t_{oss} opportuno (per $\hat{\beta}_0$ o$\hat{\beta}_1$)
4. Controllo se $\mid t_{oss} \leq t_{n-p,1-\alpha/2}$, se è vero accetto H_0, altrimenti la rifiuto.

B.3 Modello lineare con più variabili

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Minimi quadrati:

$$s^2(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i})$$

$$\begin{cases} \hat{\beta}_1 & = \frac{cov(X_1,Y)sqm(X_2)-cov(X_2,Y)cov(X_1,X_2)}{sqm(X_1)sqm(X_2)-cov^2(X_1,X_2)} \\ \hat{\beta}_2 & = \frac{cov(X_2,Y)sqm(X_1)-cov(X_1,Y)cov(X_1,X_2)}{sqm(X_1)sqm(X_2)-cov^2(X_1,X_2)} \\ \hat{\beta}_0 & = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 \end{cases}$$

Appendice C

Soluzioni Parziali

C.1 Aprile 2015

C.1.1 Esercizio 5

Domanda a

Domanda b

Bisogna completare i dati dell'output.

I gradi di libertà per la F_{oss} sono 2 e 19, rispettivamente $p-1$ e $n-p$ con p numero di parametri. I gradi di libertà degli errori sono sempre $n-p=19$.

Per calcolare RSE serve s^2 che è la stima della varianza degli errori (e dei residui).

$$RSE = \sqrt{\frac{SSE}{n-p}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}} = \sqrt{s^2}$$

Sappiamo che:

$$\widehat{\text{var}(\vec{\hat{\beta}})} = (\mathbf{X}^T\mathbf{X})^{-1}s^2$$

e che $\sqrt{\text{var}(\hat{\beta}_0)} = 126,758$ e $(\mathbf{X}^T\mathbf{X})^{-1}_{11} = 0,6363$ da cui segue $s^2 = \frac{126,758^2}{0,6363} = 25.251,596$, pertanto $RSE = \sqrt{s^2} = 158,...$