MODELLI PER DATI CROSS-SECTION

Vedremo

- cos’è l’Econometria
- il Modello di Regressione semplice
- il Modello di Regressione multipla

Che cos'è l'Econometria

Econometria = Misurazione dei fenomeni economici

Applicazione di metodi statistici e matematici per l'analisi dei dati economici, con il fine di dare validazione empirica alle teorie economiche, verificarle, oppure rigettarle (Maddala, 1992)

Un’altra definizione

But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonymous with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics

Perché l'Econometria?

- Quando si desidera vagliare la consistenza di una teoria economica con i dati
- Quando vogliamo dare una valutazione quantitativa dell'efficacia delle manovre di politica economica
- Quando vogliamo quantificare una certa relazione che ha una qualche rilevanza per le decisioni di impresa

La Nozione di modello

- Il compito principale dell'Econometria è quantificare **relazioni tra variabili economiche** sulla base dei **dati disponibili**, usando tecniche **statistiche**
- Un'analisi empirica usa i dati per sottoporre a verifica (test) una teoria o per stimare una relazione
- Nel sottoporre a verifica empirica una teoria è necessario disporre di un modello economico formale
- Gli economisti formulano modelli che descrivono le relazioni fra diverse variabili, per esempio la relazione tra reddito e consumo, oppure tra salari e stipendi percepiti e il livello di scolarità, il genere, la regione geografica, ecc.

Il Modello econometrico

- **Specificazione**: formalizzazione in termini matematico-statistici delle ipotesi teoriche in base alle informazioni empiriche a disposizione (individuazione delle variabili rilevanti per l'analisi, scelta della forma funzionale della legge sottostante il fenomeno studiato e ipotesi sugli errori commessi)
- **Stima dei parametri del modello**: quantificazione delle relazioni economiche studiate. In questa fase si cerca di individuare la struttura del modello che meglio approssima la vera struttura del fenomeno in oggetto (incognita)
- **Verifica della validità del modello**: sequenza di operazioni volte a valutare la validità del modello sulla base delle osservazioni disponibili. Comporta la verifica della validità del modello formale prescelto (attendibilità della specificazione, capacità descrittiva, conformità alle aspettative teoriche, capacità previsionale)

Prima formalizzazione

In economia si usa il concetto matematico di una funzione per esprimere delle idee sulle relazioni tra variabili economiche. Per esempio, per descrivere la relazione fra consumo e reddito si può scrivere:

$$Consumo = f\ (Reddito)$$

per indicare che il livello di consumo è dato da una qualche funzione $f\ (\cdot)$ del reddito
Oppure:

$$Q^d = f\ (P, P^s, P^c, Reddito)$$

dove il prezzo di un'automobile dipende dal prezzo della vettura P, dal prezzo di automobili sostitute P^s, dal prezzo di beni complementari P^c e dal livello di reddito

Errore nel modello

Un modello econometrico consiste in una parte sistematica e in un componente casuale e non prevedibile che chiameremo un **errore casuale**

$$Q^d = f(P, P^s, P^c, Reddito) + e$$

$$f(P, P^s, P^c, Reddito) = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 Reddito$$

$$Q^d = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 Reddito + e$$

Parametri del modello

I coefficienti $\beta_1, \beta_2, \ldots, \beta_5$ sono parametri sconosciuti del modello che stimiamo utilizzando dati economici e una tecnica econometrica

- La forma funzionale rappresenta un'ipotesi sulla relazione tra le variabili
- In ogni particolare problema, l'obiettivo consiste nell'individuare una forma funzionale compatibile con la teoria economica e i dati
- La parte sistematica è la parte che otteniamo dalla teoria economica e include un'ipotesi sulla forma funzionale
- La componente casuale rappresenta un componente *rumore* che oscura la nostra comprensione della relazione tra le variabili e che rappresentiamo usando la variabile casuale *e*

Problema

Gli economisti, di solito sono interessati a studiare le relazioni tra variabili

- La teoria economica ci dice che la spesa per beni economici dipende dal reddito
- Di conseguenza, chiamiamo *y* la **variabile dipendente** e *x* la **variabile indipendente** o **esplicativa**
- In econometria, riconosciamo che le spese nel mondo reale sono **variabili casuali** e vogliamo utilizzare i dati per conoscere la relazione

I Dati

Per studiare la relazione esistente tra reddito familiare e spesa in beni alimentari conduciamo il seguente esperimento:

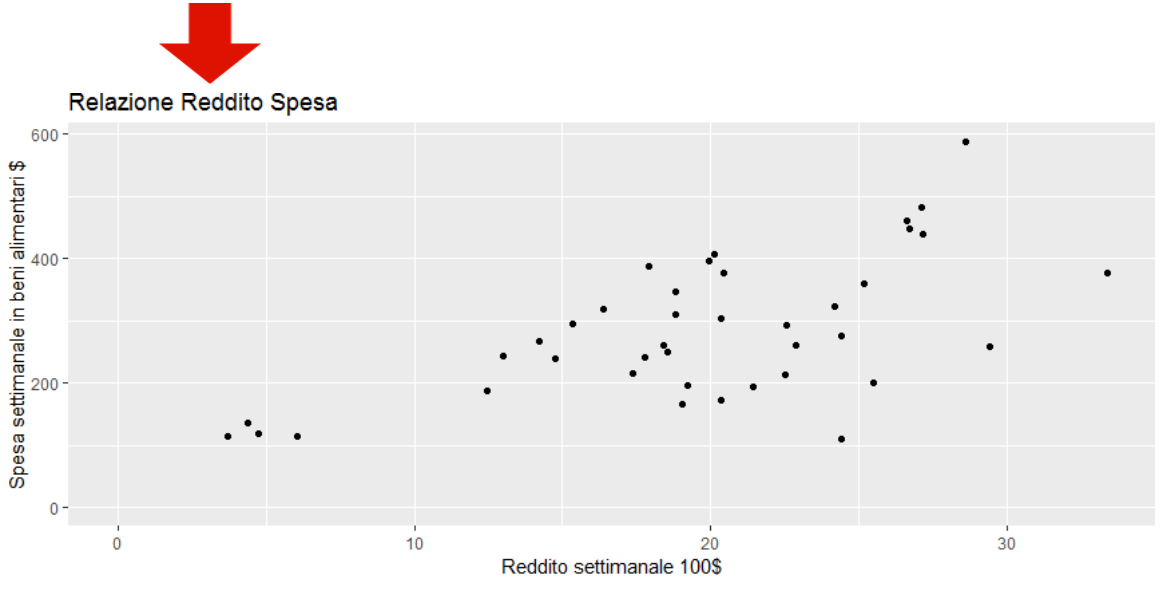
1. Definiamo la popolazione di riferimento (famiglie residenti in una certa città, regione o stato)
2. Estraiamo dalla popolazione un campione casuale di un certo numero di famiglie con reddito medio di 100 euro al mese e chiediamo "qual è stata la vostra spesa pro capite in beni alimentari la settimana scorsa?"
3. Tale spesa è la nostra variabile casuale *y* perché il suo valore è ignoto fino al momento in cui viene selezionata la famiglia

Nell'analizzare la relazione economica dobbiamo prendere atto che stiamo descrivendo il comportamento **medio** o **sistematico** di **molte** famiglie: la spesa in alimenti varierà da una famiglia all'altra per svariate ragioni (tipi di cibi che dipendono dai gusti, età media della famiglia, tipo di dieta)

La Regressione

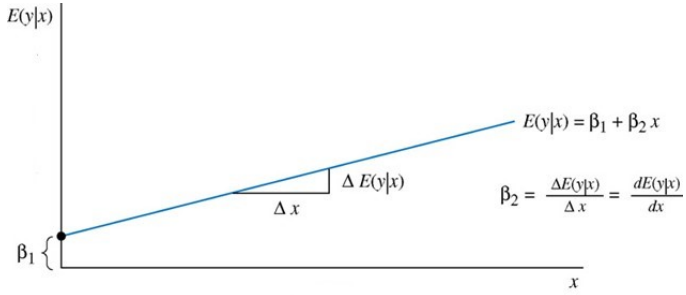
- Più in generale, ci domandiamo quale sia l'effetto (ignoto a priori) della variazione in una variabile *x* (reddito) su un'altra variabile *y* (consumi)
- Per fornire una risposta a tali domande si può utilizzare il Modello di regressione
- Prima di decidere quale Modello di regressione adottare è preferibile analizzare i dati a disposizione
  - Rappresentazione grafica
  - Stima econometrica

La Rappresentazione grafica



La Rappresentazione matematica

Funzione di regressione semplice: $E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x$



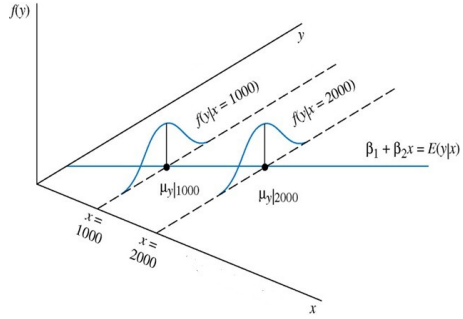
I parametri di regressione:

- β_1: media settimanale spesa per una famiglia con reddito nullo
- β_2: propensione marginale al consumo ci dice di quanto varia il reddito per una data variazione del reddito disponibile

La Rappresentazione econometrica

Per ogni livello del reddito il valore atteso, o media, della spesa familiare (y) è descritto dalla funzione di regressione

$E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x$



Prima ipotesi: dal grafico risulta che la distribuzione della spesa familiare è ipotizzata essere centrata sulla media $E(y|x)$ qualunque sia il livello del reddito

Errore

- La relazione lineare non vale con esattezza: le discrepanze tra valori osservati di *y* e quelli derivanti da una relazione esatta con *x* possono dipendere da **errori di specificazione** (altre variabili esplicative non considerate nel modello) o **errori di misura** presenti nella variabile *y*
- L'essenza dell'analisi di regressione è che ogni osservazione della variabile dipendente può essere scomposta in due parti: una **componente sistematica** e una **componente casuale**

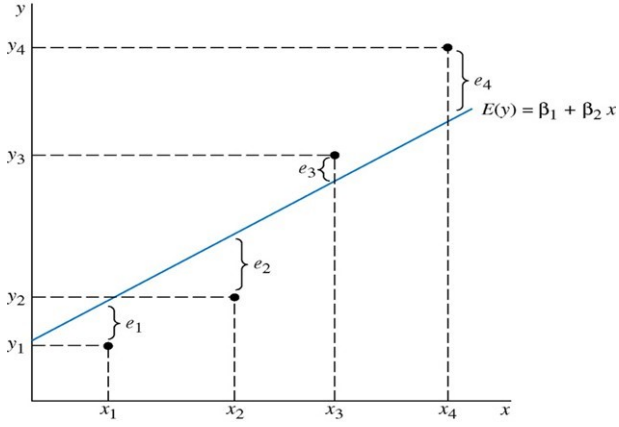
$$e = y - E(y|x) = y - \beta_1 - \beta_2 x$$

$$y = \beta_1 + \beta_2 x + e$$

Sia *e* che *y* sono variabili casuali: *y* è osservabile mentre *e* è non osservabile

Errore e Retta di regressione

Se i parametri di regressione fossero noti, *y* potrebbe essere decomposta nelle due componenti



La Spesa alimentare: i dati

Tabella: Spesa e reddito delle famiglie

Famiglia	Spesa Alimentare	Reddito settimanale (100 $)
1	115.22	3.69
2	135.98	4.39
.	.	.
.	.	.
39	257.95	29.4
40	375.73	33.4

L'econometria affronta il problema di usare i dati campionari (x_i, y_i) per ottenere le stime numeriche dei parametri incogniti β_1 e β_2

La Stima dei minimi quadrati

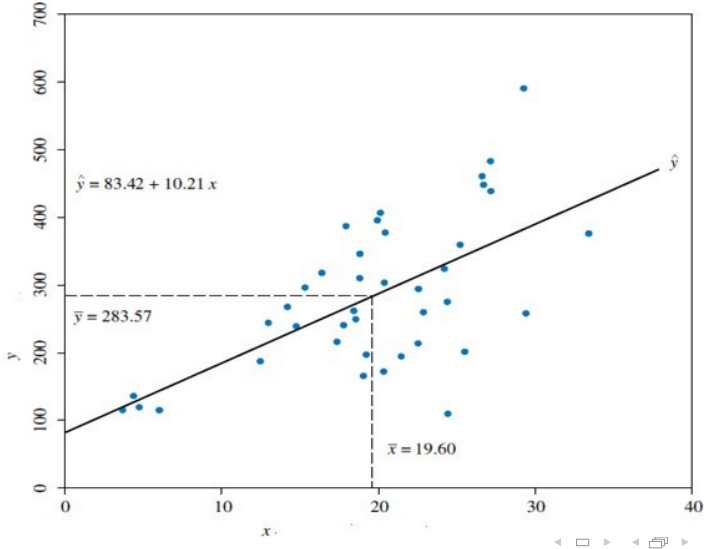
- Come possiamo scegliere tra le tante Rette di regressione?
- Ipotesi sulla componente di errore
- Il processo di stima più utilizzato è il metodo dei mimimi quadrati ordinari (OLS, dall'inglese Ordinary Least Squares)
- Il termine minimi quadrati deriva dal fatto che la retta ha la proprietà di minimizzare la somma dei quadrati delle distanze dei punti dalla retta stessa

$$S(\beta_1, \beta_2) = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2$$

La Stima dei minimi quadrati in R

```
> data(food)
> str(food)
'data.frame': 40 obs. of  2 variables:
 $ food_exp: num  115 136 119 115 187 ...
 $ income  : num  3.69 4.39 4.75 6.03 12.47 ...
> mod1 <- lm(food_exp~income, data=food)
```

Spesa alimentare: la Stima della Retta di regressione



Verifica d'ipotesi: in generale

- A partire da un modello economico e statistico, vengono formulate delle **ipotesi** sul comportamento economico
- Queste ipotesi sono quindi rappresentate come affermazioni sui parametri del modello
- Le verifiche di ipotesi usano le informazioni sui parametri contenuta nel campione di dati, rappresentata dalle stime puntuali dei minimi quadrati e i rispettivi standard error, per giungere ad una conclusione sulla validità delle ipotesi

Verifica d'ipotesi: stima minimi quadrati

- L'ipotesi nulla è espressa come $H_0 : \beta_k = c \quad k = 1, 2$, dove c è una costante ed è un valore importante nel contesto di un modello di regressione specifico (un valore comune per c è 0)
- L'ipotesi alternativa H_1, flessibile e che dipende in una certa misura dalla teoria economica, viene accettata se l'ipotesi nulla viene rifiutata (tuttavia considereremo ipotesi alternativa $\beta_k \neq c$)
- Sulla base del valore di una statistica test decidiamo di rifiutare l'ipotesi nulla o di non rifiutarla. Una statistica test ha una caratteristica speciale: la sua distribuzione di probabilità è completamente nota quando l'ipotesi nulla è vera, e ha qualche altra distribuzione se l'ipotesi nulla non è vera

$$t = \frac{b_k - c}{se(b_k)} \sim t_{N-2}$$

Verifica d'ipotesi: p-value in R

- Quando si illustrano i risultati di un test statistico di ipotesi è diventata pratica standard riportare il **p-value** (un'abbreviazione probability value) del test (livello di significativit`a empirico). Conoscendo il valore p del p-value di un test, p, possiamo determinare l'esito del test confrontando il valore p con il livello di significativit`a scelto, α, senza cercare o calcolare i valori critici
- Nel nostro caso si rifiuta l'ipotesi nulla quando il valore p è inferiore o uguale al livello di significatività α. Cioè, se $p \leq \alpha$ allora si rifiuta H_0. Se $p > \alpha$ allora non è possibile rifiutare H_0

```
> mod1 <- lm(food_exp~income, data=food)
> summary(mod1)
```

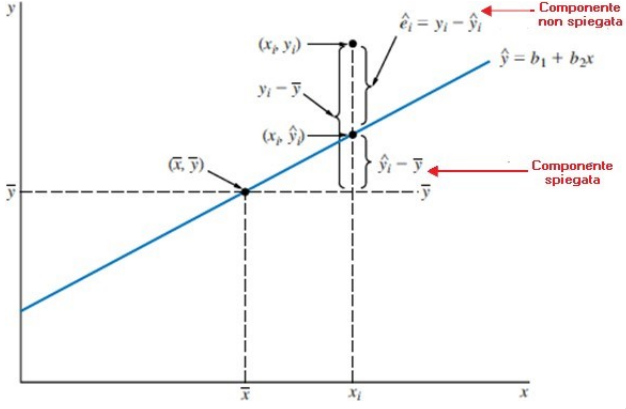
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    83.416     43.410   1.922   0.0622 .
food$income    10.210      2.093   4.877 1.95e-05 ***
```

Una Misura di adattamento del modello: R^2

$y_i = \hat{y}_i + \hat{e}_i$

$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{e}_i$

$y_i - \bar{y} =$ comp. spiegata dal modello + comp. non spiegata dal modello



Coefficiente di determinazione R^2

$$R^2 = \frac{SQM}{SQT} = 1 - \frac{SQR}{SQT}$$

- **SQT**: somma dei quadrati totale che misura la variazione totale in *y* attorno alla media campionaria
- **SQM**: somma dei quadrati dovuta al modello di regressione che misura la parte della variazione totale in *y* attorno alla media campionaria spiegata dal modello
- **SQR**: somma dei quadrati dei residui che riflette parte della variazione in *y* non spiegata dal modello
- Quanto l'R^2 è più vicino a 1, tanto piu` i valori del campione y_i sono vicini all'equazione di regressione adattata. Se $R^2 = 1$, tutti i punti campionari sono esattamente sulla linea dei minimi quadrati, quindi $SQR = 0$, e il modello si adatta ai dati "perfettamente". Se i dati per *y* e *x* non sono correlati e non mostrano alcuna associazione lineare, allora la stima della retta di regressione `e orizzontale e identica a *y*, quindi $SQM = 0$ e $R^2 = 0$

Coefficiente di determinazione R^2 in R

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    83.416     43.410   1.922   0.0622 .
food$income    10.210      2.093   4.877 1.95e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.52 on 38 degrees of freedom
Multiple R-squared:  0.385,     Adjusted R-squared:  0.3688
```

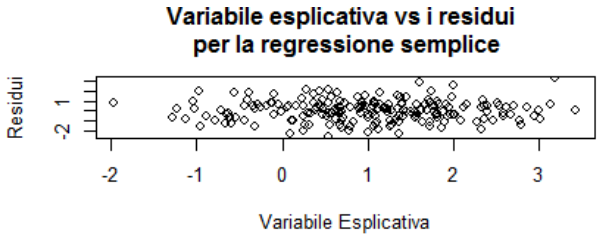
Concludiamo che il 38,5% della variazione nella spesa alimentare (rispetto alla sua media campionaria) è spiegato dal nostro modello di regressione, che utilizza solo il reddito come variabile esplicativa.

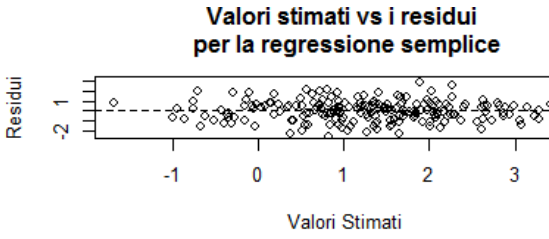
Per valutare il modello è altrettanto importante considerare il segno e la grandezza delle stime, la loro significatività statistica ed economica, la loro precisione e la capacità del modello stimato di prevedere valori della variabile dipendente che non appartengono al campione di stima

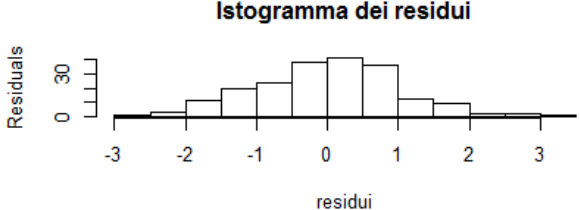
Analisi grafica dei residui

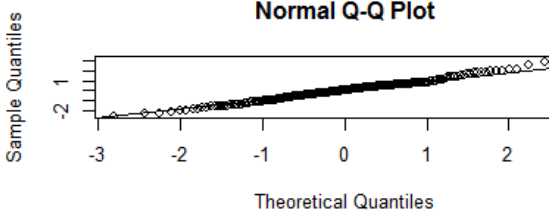
- **Grafico variabili esplicative verso i residui:** permette di individuare non corrette specificazioni della dipendenza dalle variabili esplicative, come ad esempio, dipendenze non lineari
- **Grafico valori stimati dal modello verso i residui:** ci permette di verificare se le ipotesi di omoschedasticità,media nulla e non correlazione dei residui sono verificate
- **Normal probability plot:** confronto tra i quantili della distribuzione dei residui osservati e quella di una normale standardizzata

Corretta specificazione









Un Modello di Regressione lineare multipla

$$Q^d = f\ (P, P^s, P^c, Reddito) + e$$

$$f\ (P, P^s, P^c, Reddito) = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 Reddito$$

$$Q^d = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 Reddito + e$$

Esempio: due variabili esplicative

Costruiamo un modello economico in cui i ricavi dipendono da una o più variabili esplicative. Inizialmente ipotizziamo che i ricavi siano linearmente correlati al prezzo e alla spesa pubblicitaria [data(andy) in library(PoEdata)]. Il modello economico è:

$$\text{Sales} = \beta_0 + \beta_1 \text{Price} + \beta_2 \text{Advert}$$

- β_1 è la variazione dei ricavi (in migliaia di dollari) quando il prezzo (Price) è aumentato di un'unità (in dollari) e la spesa pubblicitaria (Advert) è tenuta costante
- B_2 è la variazione ricavi (in migliaia di dollari) quando la spesa pubblicitaria (Advert) è aumentata di un'unità (in migliaia di dollari) e il prezzo (Price) è tenuto costante

Stima e test di ipotesi: Parametri

Interpretazione della t di Student rimane la stessa

```
> summary(mod1)

Call:
lm(formula = sales ~ price + advert, data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-13.4825  -3.1434  -0.3456   2.8754  11.3049

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.9136     6.3516  18.722  < 2e-16 ***
price        -7.9079     1.0960  -7.215 4.42e-10 ***
advert        1.8626     0.6832   2.726  0.00804 **
```

- **Il coefficiente negativo di PRICE suggerisce che la domanda è elastica:** stimiamo che, con la pubblicità mantenuta costante, un aumento del prezzo di 1$ porterà un calo delle entrate mensili di $7,908$
- **Il coefficiente di Advert è positivo:** stimiamo che con il prezzo costante, un aumento della spesa pubblicitaria di $1,000$ porterà ad un aumento del fatturato di $1,863$

Test d'Ipotesi congiunta

Il test per verificare un'Ipotesi nulla congiunta utilizza la statistica test F calcolata usando una semplice formula basata sulla somma dei quadrati dei residui di due regressioni:

- nella prima regressione, chiamata regressione vincolata (V), l'ipotesi nulla assume che i coefficienti siano zero escludendo, quindi, i regressori rilevanti dalla regressione
- nella seconda regressione, chiamata regressione non vincolata (NV), `e valida invece l'ipotesi alternativa

Se la somma dei quadrati dei residui `e sufficientemente più piccola nella regressione non vincolata rispetto alla regressione vincolata, il test rifiuta l'ipotesi nulla

$$F = \frac{(SQR_V - SQR_{NV})/q}{SQR_{NV}/(n-k-1)} \sim F_{q,n-k-1}$$

Verifica complessiva del modello

Consideriamo la nostra regressione quadratica

Modello per i ricavi

$$Sales = \beta_0 + \beta_1 Price + \beta_2 Advert + \beta_3 Advert^2 + \epsilon_i \quad i = 1, \ldots, n$$

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

H_1 : Almeno uno dei β_k = 1, 2, 3

Statistica F in R lm()

```
Call:
lm(formula = sales ~ price + advert + I(advert^2), data = andy)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2553  -3.1430  -0.0117   2.8513  11.8050

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.7190     6.7990  16.137  < 2e-16 ***
price        -7.6400     1.0459  -7.304 3.24e-10 ***
advert       12.1512     3.5562   3.417  0.00105 **
I(advert^2)  -2.7680     0.9406  -2.943  0.00439 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.645 on 71 degrees of freedom
Multiple R-squared:  0.5082,    Adjusted R-squared:  0.4875
F-statistic: 24.46 on 3 and 71 DF,  p-value: 5.6e-11
```

Scelta del modello

- Gli strumenti principali per costruire un modello dovrebbero essere la teoria economica, un valido ragionamento basato su princìpi economici e fare in modo che il modello soddisfi le ipotesi sugli errori
- Si dovrebbe anche considerare la possibilità di distorsioni variabili omesse e l'esclusione di variabili irrilevanti che possono aumentare la variabilità delle stime
- Dopo aver considerato tutti questi aspetti e stabilito un modello, ci sono alcune grandezze che aiutano a confrontare diversi modelli oltre a R^2, R^2 aggiustato (R^{-2}), il criterio di informazione di Akaike (AIC), e il criterio di Schwarz (o Bayesian information) (SC o BIC)
- Il criterio di informazione di Akaike (AIC) e il criterio di Schwarz usano la stessa idea dell'R^2 di penalizzare l'introduzione di regressori extra e il miglior modello è quello che minimizza i criteri

Multicollinearità

- **Perfetta multicollinearità** se uno dei regressori è una esatta combinazione lineare degli altri regressori. A seconda dei software, il pacchetto gestisce la perfetta multicollinearità in due modi: ometterà una delle variabili o rifiuterà di calcolare le stime OLS e darà un messaggio di errore
- La **multicollinearità imperfetta** sorge quando uno dei regressori è altamente correlato, anche se non perfettamente con gli altri regressori. A differenza della perfetta multicollinearità, la multicollinearità imperfetta non impedisce la stima della regressione. Tuttavia, implica che uno o più coefficienti della regressione possa essere stimato in modo impreciso

Riepilogo e conclusioni finali

- cos'è l'Econometria
- il Modello di Regressione semplice
- il Modello di Regressione multipla