Analisi della regressione multipla

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$

- 2. Inferenza

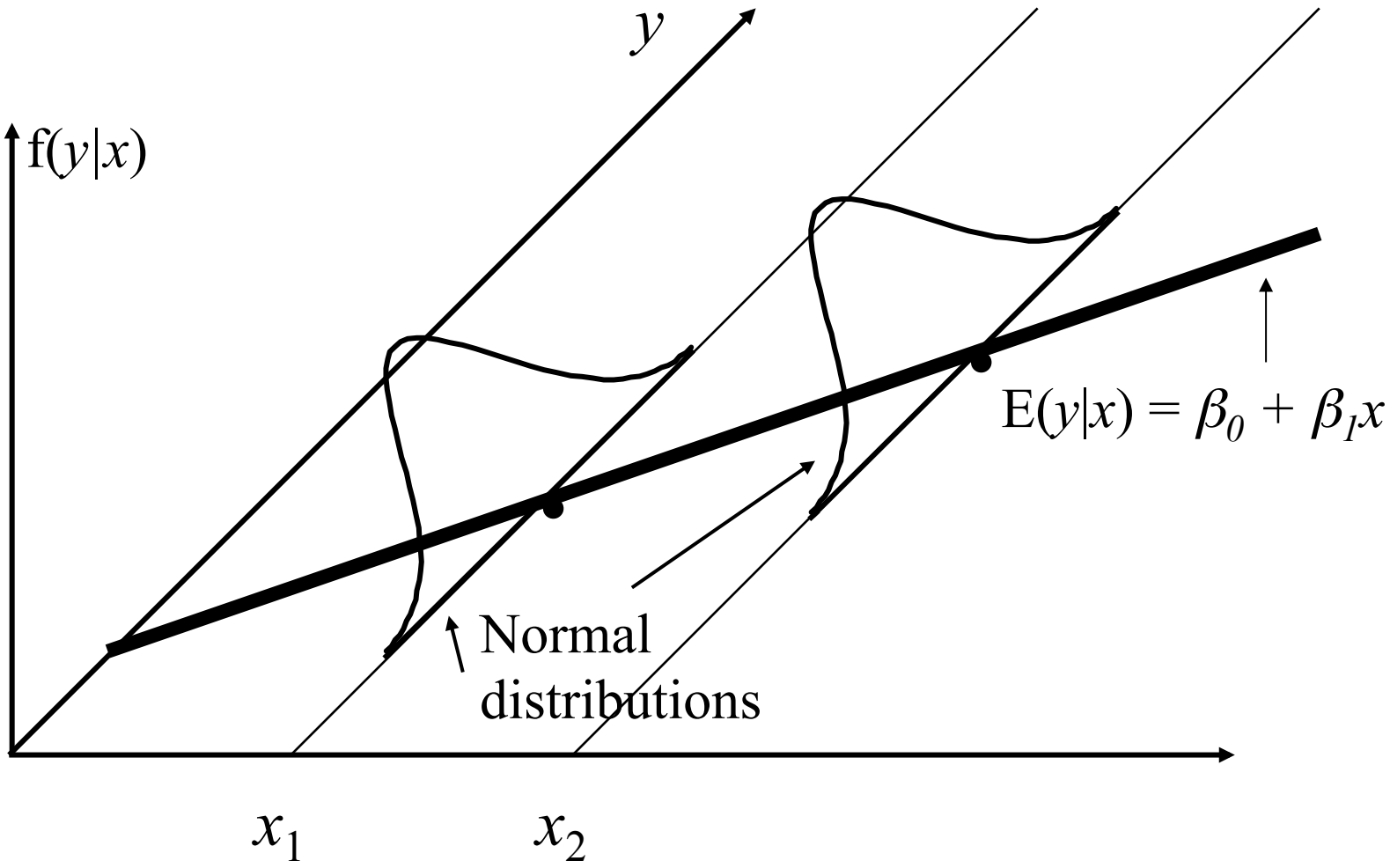
Assunzione del Modello Classico di Regressione Lineare (CLM)

- Sappiamo che, date le assunzioni Gauss-Markov, OLS è BLUE,
- Per effettuare la verifica delle ipotesi secondo l'approccio classico, abbiamo bisogno di un'altra assuzione (oltre le assunzioni Gauss-Markov)
- Bisogna assumere che u è indipendente dalle $x_1, x_2, \ldots, x_k$ e u è distribuita normalmente con media zero e varianza σ^2: $u \sim \text{Normal}(0,\sigma^2)$

Assunzioni CLM (cont)

- Se sono valide le assunzioni CLM, OLS è non solo BLUE, ma ha anche la minima varianza tra gli stimatori lineari corretti
- E' possibile sintetizzare le assunzioni CLM nel seguente modo
- $y|\boldsymbol{x} \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$
- Per il momento assumiamo una distribuzione normale, chiaramente non è sempre questo il caso
- Nei grandi campioni l'assunzione di normalità può essere rilassata

Distribuzione normale omoschedastica con una sola Variabile esplicativa



Distribuzione campionaria normale

Date le assunzioni CLM, condizionati ai valori campionari delle variabili indipendenti

$\hat{\beta}_j \sim \text{Normal}\left[\beta_j, Var\left(\hat{\beta}_j\right)\right]$, tale che

$$\left(\hat{\beta}_j - \beta_j\right) \Big/ sd\left(\hat{\beta}_j\right) \sim \text{Normal}(0,1)$$

$\hat{\beta}_j$ è distribuito normalmente poiché è dato da una combinazione lineare degli errori

Il Test *t*

Date le assunzioni CLM

$$\left(\hat{\beta}_{\mathrm{j}} - \beta_j\right) \Big/ se\left(\hat{\beta}_j\right) \sim t_{n-k-1}$$

Notare che si tratta di una "*t* distribution" (vs normal)

dato che dobbiamo stimare σ^2 attraverso $\hat{\sigma}^2$

I gradi di libertà sono:$n-k-1$

Il *t* Test (cont)

- Sapendo che la distribuzione campionaria dello stimatori ci consente di effettuare la verifica delle ipotesi
- Specificare l'ipotesi nulla
- Per esempio, H_0: $\beta_j=0$
- Se non rifiutiamo l'ipotesi nulla, allora non si rifiuta che x_j non ha effetti su y, controllando per le altre x

Il *t* Test (cont)

Per calcolare il test dobbiamo determinare

"la statistica t" per $\hat{\beta}_j$: $t_{\hat{\beta}_j} \equiv \hat{\beta}_j \Big/ se\left(\hat{\beta}_j\right)$

Utilizzeremo la statistica t insieme alla regola di rifiuto per determinare se rifiutare o non rifiutare l'ipotesi nulla, H_0

t Test: Ipotesi Alternativa ad una coda

- Oltre all'ipotesi nulla, H_0, dobbiamo specificare un'ipotesi alternativa, H_1, e un livello di significatività
- H_1 può essere ad una coda o a due code
- H_1: $\beta_j > 0$ e H_1: $\beta_j < 0$ è ad una coda
- H_1: $\beta_j \neq 0$ è a due code
- Se vogliamo una probabilità del 5% di rifiutare H_0 quando è vera, diciamo allora che il livello di significatività è pari al 5%

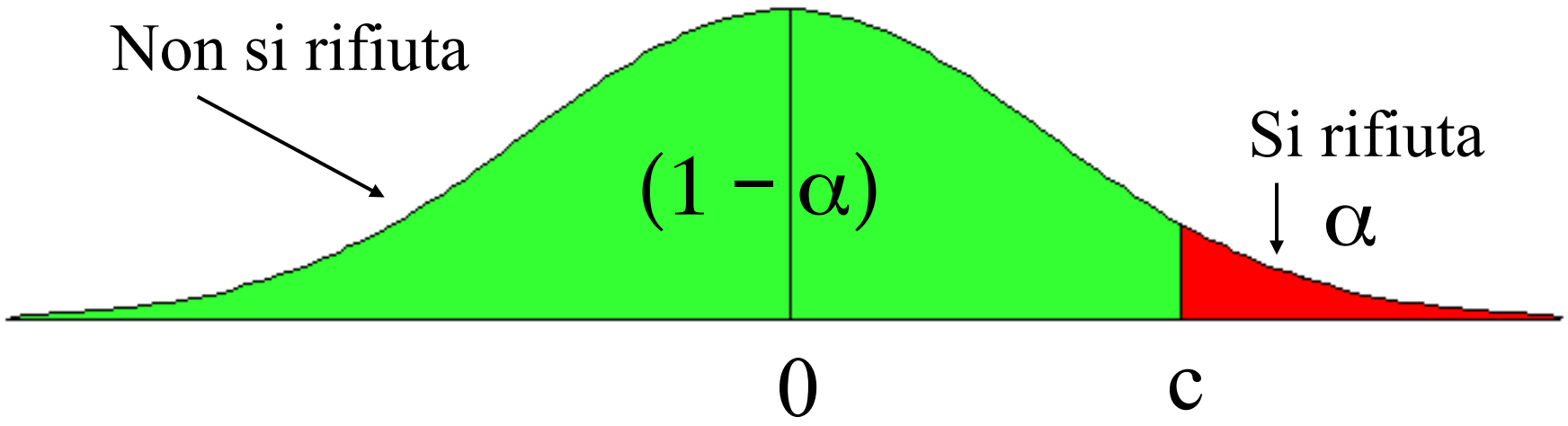
t Test: Ipotesi Alternativa ad una coda (cont.)

- Dopo aver fissato il livello di significatività, α, valutiamo il percentile $(1 - \alpha)^{th}$ nella distribuzione *t* con $n - k - 1$ "df" (gradi di libertà) e lo definiamo *c*, valore critico
- Si rifiuta l'ipotesi nulla se la statistica *t* è maggiore del valore critico
- Se la statistica *t* è minore del valore critico non rifiutiamo l'ipotesi nulla

t Test: Ipotesi Alternativa ad una coda (cont.)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

$H_0: \beta_j = 0$ $H_1: \beta_j > 0$



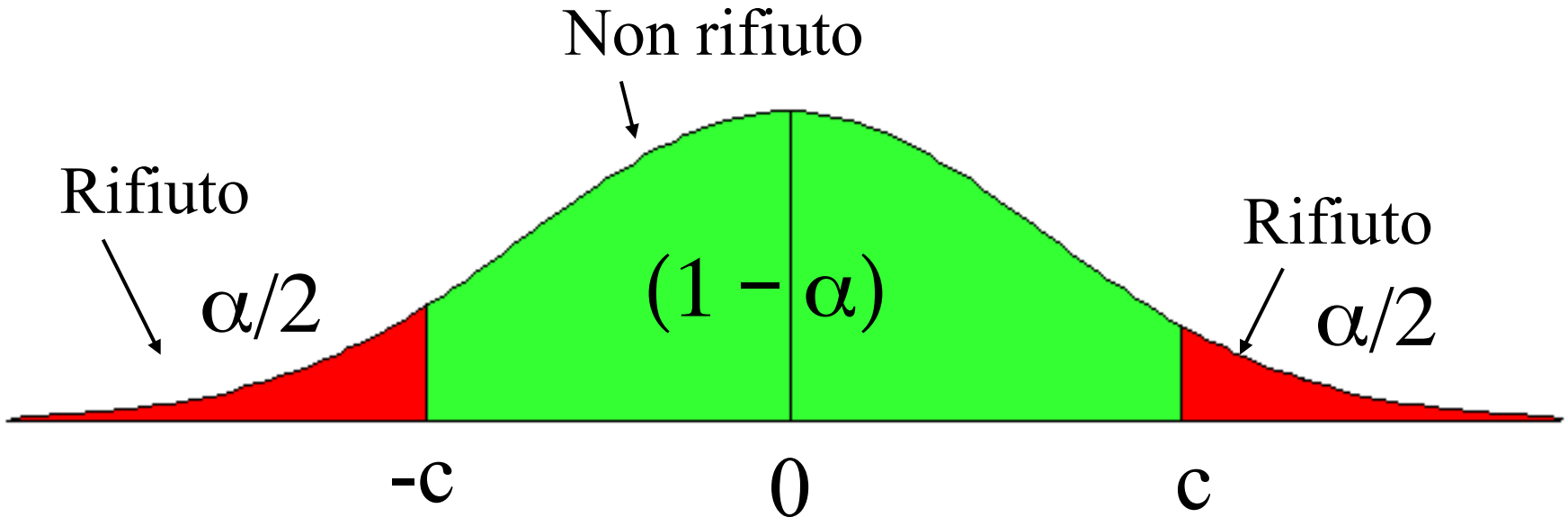
Ipotesi alternativa ad una coda vs due code

- Dato che la distribuzione *t* è simmetrica, risulta semplice testare H_1: $\beta_j < 0$. Dobbiamo soffermarci sulla parte negativa della distribuzione e guardare il valore critico negativo
- Si rifiuta l'ipotesi nulla se la statistica $t < -c$, e non si può rifiutare l'ipotesi nulla se la statistica $t >$ di $-c$
- Nel caso di ipotesi alternativa a due code, bisogna stabilire il valore critico basato su $\alpha/2$ e rifiutare H_1: $\beta_j \neq 0$ se il <u>valore assoluto</u> della statistica $t > c$

Ipotesi alternativa a due code

$$y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + u_i$$

$H_0: \beta_j = 0$ $\qquad$ $H_1: \beta_j > 0$



Sintesi H_0: $\beta_j = 0$

- Se l'ipotesi alternativa non è specificata, si assume essere a due code
- Se rifiutiamo l'ipotesi nulla, si conclude che "x_j è statisticamente significativa al livello del α %"
- Se non possiamo rifiutare l'ipotesi nula, concludiamo che "x_j è statisticamente non significativa al livello di α %"

Test per altre Ipotesi

- Una formulazione più generica della statistica *t* ci permette di verificare l'ipotesi del tipo H_0: $\beta_j = a_j$
- In questo caso, la statistica *t* appropriata è

$$t = \left(\hat{\beta}_j - a_j\right) \Big/ se\left(\hat{\beta}_j\right), \text{ dove}$$

$a_j = 0$ nel caso del test consueto

Intervalli di Confidenza

- Un altro modo per utilizzare l'inferenza classica consiste nel costruire gli intervalli di confidenza usando gli stessi valori critici utilizzati nel caso del test a due code
- Un intervallo di confidenza (1 - α) % è definito:

$$\hat{\beta}_j \pm c \bullet se\left(\hat{\beta}_j\right), \text{ dove c è il } \left(1-\frac{\alpha}{2}\right) \text{ percentile}$$

$$\text{nella distribuzione } t_{n-k-1}$$

Calcolo del *p*-value per il test *t*

- Un metodo alternativo all'approccio classico discusso consiste nel chiedersi, "quale è il valore minore del livello di significatività che comporta il rifiuto dell'ipotesi nulla?"
- Quindi, bisogna calcolare la statisitca *t*, e valutare quali percentili sono nella corretta distribuzione *t* – questo rappresenta il *p*-value
- *p*-value rappresenta **la probabilità di osservare la statistica *t* calcolata, nel caso l'ipotesi nulla fosse vera**

Software econometrici, *p*-values, *t* tests, ecc.

- La maggior parte dei software econometrici calcolano il *p*-value assumendo un test a due code
- Se siete interessati ad un'ipotesi alternativa ad una coda, bisogna dividere il *p*-value per 2
- Excel calcola la statistica *t*, *p*-value, e l'intervallo di confidenza di 95% relativo a H_0: $\beta_j = 0$, nelle colonne indicate con "t", "P > |t|" e "[95% Conf. Interval]". Gretl calcola la statistica *t*, *p*-value.

Test di Combinazioni Lineari

- Supporre di voler testare non se β_1 è uguale ad una costante, ma se è uguale ad un altro, cioé $H_0 : \beta_1 = \beta_2$
- Si usa la stessa procedura per il calcolo della statistica t

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se\left(\hat{\beta}_1 - \hat{\beta}_2\right)}$$

Test di Combinazioni Lineari (Cont.)

Dato

$se\left(\hat{\beta}_1 - \hat{\beta}_2\right) = \sqrt{Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)}$, allora

$$Var\left(\hat{\beta}_1 - \hat{\beta}_2\right) = Var\left(\hat{\beta}_1\right) + Var\left(\hat{\beta}_2\right) - 2Cov\left(\hat{\beta}_1, \hat{\beta}_2\right)$$

$$se\left(\hat{\beta}_1 - \hat{\beta}_2\right) = \left\{\left[se\left(\hat{\beta}_1\right)\right]^2 + \left[se\left(\hat{\beta}_2\right)\right]^2 - 2s_{12}\right\}^{1/2}$$

dove s_{12} è la stima di $Cov\left(\hat{\beta}_1, \hat{\beta}_2\right)$

Test di Combinazioni Lineari (Cont.)

- Per il calcolo della formula specificata, abbiamo bisogno di s_{12}, che non viene fornito automaticamente dai sofware econometrici (Gretl, Excel)
- Alcuni software hanno un'opzione che permette il calcolo di s_{12}
- In Gretl, Dopo aver effettuato la regressione y x1 x2 … xk bisogna scegliere l'opzione "test"e successivamente "restrizioni lineari" e scrivere b1 - b2 = 0 (dove b1 e b2 sono i coefficienti di x1 e x2) e si ottiene la statistica t e il *p*-value del test
- In generale, si può riscrivere il modello econometrico per testare l'ipotesi sulle restrizioni lineari

Esempio:

- Supporre di essere interessati agli effetti della pubblicità elettorale sui risultati delle elezioni
- Il modello è $voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u$
- H_0: $\beta_1 = -\beta_2$, o H_0: $\theta_1 = \beta_1 + \beta_2 = 0$
- $\beta_1 = \theta_1 - \beta_2$, con sostituzioni e riarrangiamenti otteniamo $\Rightarrow$ $voteA = \beta_0 + \theta_1 \log(expendA) + \beta_2 \log(expendB - expendA) + \beta_3 prtystrA + u$

Esempio (cont):

- Questo rappresenta lo stesso modello di prima con l'unica differenza che ora otteniamo dirretamente come output di regressione del software la deviazione standard per $\beta_1 - \beta_2 = \theta_1$
- Ogni combinazione lineare dei parametri può essere testata in maniera simile
- Altri esempi di una singola combinazione lineare dei parametri:
  - $\beta_1 = 1 + \beta_2$; $\beta_1 = 5\beta_2$; $\beta_1 = -1/2\beta_2$; ecc

Restrizioni Lineari Multiple

- Al momento ci siamo limitati a verificare una singola restrizione lineare, (per esempio $\beta_1 = 0$ o $\beta_1 = \beta_2$)
- Tuttavia, è possibile che siamo interessati a verificare una serie di ipotesi sui parametri
- Un tipico esempio consiste nel testare le "restrizioni di esclusione" – ovvero vogliamo sapere se un gruppo di parametri congiuntamente sia uguale a zero

Test delle Restrizioni di Esclusione

- L'ipotesi nulla potrebbe essere del tipo H_0: $\beta_{k-q+1} = 0, \ldots, \beta_k = 0$
- L'alternativa semplicemente H_1: H_0 non vera
- Non possiamo controllare le statistiche *t* separatamente per ogni parametro, dato che interessa sapere se *q* parametri **congiuntamente** sono significativi ad un dato livello di confidenza – può capitare che nessuno invece sia statisticamente significativo individualmente

Test delle Restrizioni di Esclusione (cont.)

- Per effettuare il test è necessario stimare il "modello ristretto" che non include le $x_{k-q+1}, \ldots, x_k$, e il "modello non ristretto" che include tutte le x
- Intuitivamente, Se la variazione in SSR è grande abbastanza da richiedere l'inclusione di $x_{k-q+1}, \ldots, x_k$

$$F \equiv \frac{\left(SSR_r - SSR_{ur}\right)/q}{SSR_{ur}/\left(n-k-1\right)}, \text{ dove}$$

r indica ristretto e

ur indica non ristretto

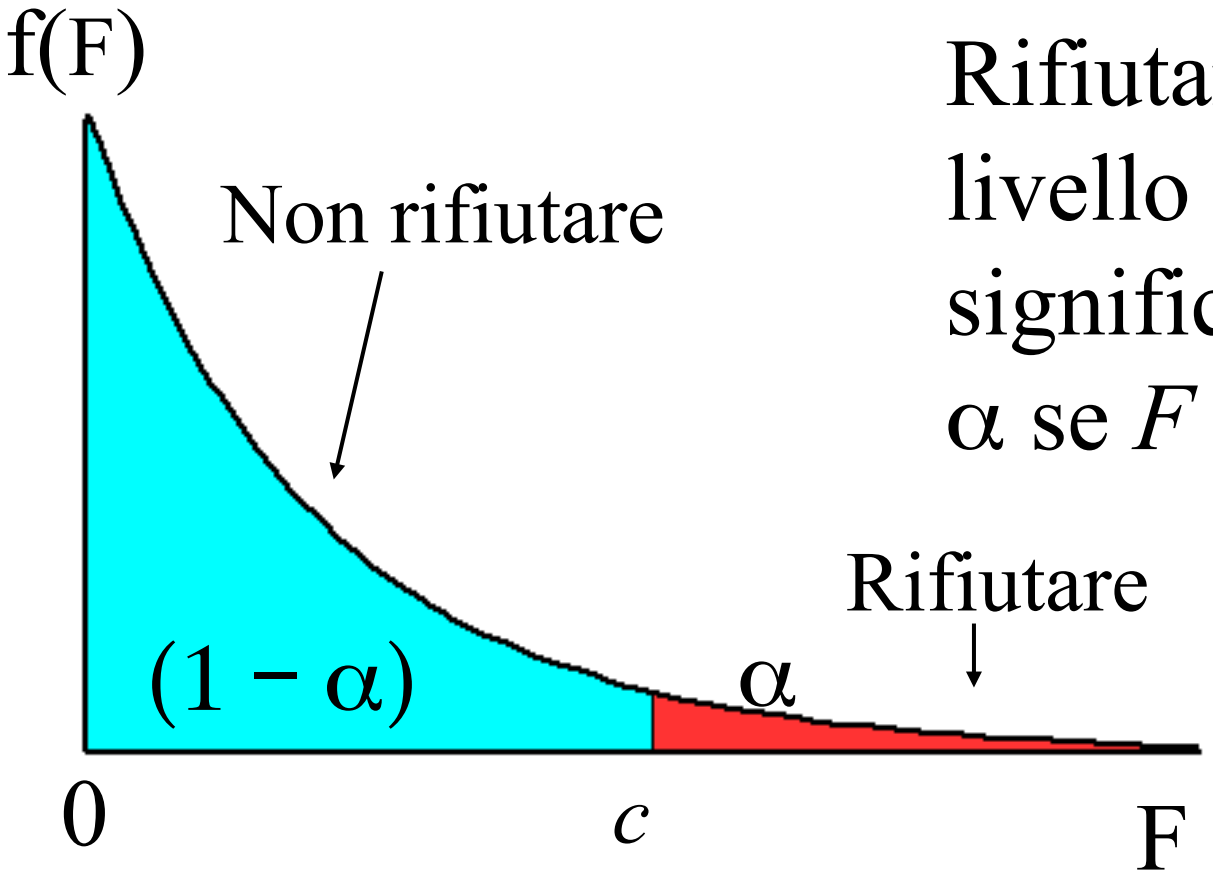
La statistica *F*

- La statistica *F* è sempre positiva, dato che il valore di SSR dal modello ristretto non può essere inferiore al valore di SSR dal modello non ristretto
- Essenzialmente la statistica *F* misura l'aumento relativo in SSR quando si passa da un modello non ristretto a un modello ristretto
- q = numero di restrizioni, o $df_r - df_{ur}$
- $n - k - 1 = df_{ur}$

La statistica *F (cont.)*

- Per decidere se l'aumento di SSR quando si passa al modello ristretto è "abbastanza elevato" da rifiutare le restrizione, abbiamo bisogno di conoscere la distribuzione campionaria della statistica *F*
- Non sorprendentemente questa è pari a, $F \sim F_{q,n-k-1}$, dove *q* si riferisce al numero dei gradi di libertà del numeratore e $n - k - 1$ al numero dei gradi di libertà del denominatore

La statistica *F (cont.)*



Rifiutare H_0 al livello di significatività di α se $F > c$

La forma R^2 della statistica F

- Poiché il valore di SSR potrebbe essere molto grande e poco trattabile si utilizza una formula alternativa
- Utilizzando l'uguaglianza di regressione SSR = SST(1 – R^2), la sostituiamo per SSR_u e SSR_{ur}

$$F \equiv \frac{\left(R_{ur}^2 - R_r^2\right)/q}{\left(1 - R_{ur}^2\right)/\left(n - k - 1\right)}, \text{ dove nuovamente}$$

r indica ristretto e ur indica non ristretto

Significatività Complessiva

- Un caso speciale delle restrizioni di esclusioni è rappresentato da H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
- Dato che R^2 di un modello con solo l'intercetta sarà pari a zero, la statistica *F* è semplicemente

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

Restrizioni Lineari

- La formula di base della statistica *F* è valida per qualsiasi restrizione lineare
- Prima bisogna stimare il modello non ristretto e poi quello ristretto
- Calcolare SSR
- Imporre le restrizioni potrebbe non essere semplice – probabilmente sarà necessario riformulare le variabili

Esempio:

- Utilizzando il modello di voto come prima
- Il modello è *voteA* = β_0 + $\beta_1$log(*expendA*) + $\beta_2$log(*expendB*) + β_3*prtystrA* + *u*
- l'ipotesi nulla H_0: $\beta_1 = 1$, $\beta_3 = 0$
- Sostituendo le restrizioni nel odello: *voteA* = β_0 + log(*expendA*) + $\beta_2$log(*expendB*) + *u*, so
- Usare *voteA* - log(*expendA*) = β_0 + $\beta_2$log(*expendB*) + *u* come modello ristretto

Statistica F Sintesi

- Nello stesso modo della statistica *t*, p-value può essere calcolato attraverso la percentuale nella appropriata distribuzione *F*
- Gretl calcola questo valore utilizzando l'opzione restrizioni lineari; fprob(*q, n* – *k* – 1, *F*), dove I valori appropriati delle distribuzioni *F, q,e n* – *k* – 1 sono utilizzati
- Nel caso si testa solo una restrizione di sclusione, allora $F = t^2$, e il valore *p*-value sarà lo stesso della distribuzione t.