Metodi Quantitativi per Economia, Finanza e Management

Lezione n°9

Regressione lineare multipla: la valutazione del modello, multicollinearità, metodi automatici di selezione dei regressor, analisi di influenza.

Il modello di regressione lineare

1. Introduzione ai modelli di regressione – Case Study
2. Obiettivi
3. Le ipotesi del modello
4. La stima del modello
5. La valutazione del modello
   - Indicatori di ‘bontà’
6. Commenti

Il modello di regressione lineare

Indicatori di 'bontà' del modello

Indicatori sintetici di bontà del Modello

- Test F → OK p-value con valori bassi
- R-quadro → OK valori alti $$R^2 = \frac{SSM}{SST}$$
- R-quadro adjusted → OK valori alti $$AdjR^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Il modello di regressione lineare

Indicatori di 'bontà' del modello

Test F per valutare la significatività congiunta dei coefficienti

- ipotesi nulla $$H_0 : \beta_1 = ... = \beta_p = 0$$

- statistica test $$F = \frac{SSM / p}{SSE / n - p - 1} \sim F(p, n - p - 1)$$

- valutazione → se p-value piccolo (rifiuto l'hp di coefficienti tutti nulli) il modello ha buona capacità esplicativa

Il modello di regressione lineare

Indicatori di 'bontà' del modello

Scomposizione della varianza SST=SSE+SSM

- total sum of squares

→ variabilità di Y

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2$$

- error sum of squares

→ variabilità dei residui

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

- model sum of squares

→ variabilità spiegata

$$SSM = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} \right)^2$$

Il modello di regressione lineare

Indicatori di 'bontà' del modello

R-quadro= SSM/SST

→ misura la % di variabilità di Y spiegata dal modello = capacità esplicativa del modello

→misura la variabilità delle osservazioni intorno alla 'retta' di regressione.

→SSM=0 (R-quadro=0) il modello non spiega

→SSM=SST (R-quadro=1) OK

- R-quadro adjusted= [1-(1-SSM/SST)] / [(n-1)(n-p-1)]

→come R-quadro ma indipendente dal numero di regressori

→ combina adattabilità e parsimonia

Il modello di regressione lineare

Indicatori di 'bontà' del modello

Indicatori sintetici di bontà del Modello

- R-quadro → Ha valori compresi tra 0 e 1
  - R-quadro = 0 => Il Modelllo non è eplicativo
  - R-quadro =1 => Il Modello spiega perfettamente
  - R-quadro >0.2/0.3 => Il Modello ha capacità esplicativa

$$R^2 = \frac{SSM}{SST}$$

- R-quadro adjusted
  - Varia tra 0 e 1
  - Ok x valori > 0.2/0.3

$$AdjR^2 = 1-(1-R^2)\frac{n-1}{n-p-1}$$

Il modello di regressione lineare

Indicatori di 'bontà' dei singoli regressori

Test t per valutare la significatività dei singoli coefficienti

- ipotesi nulla (j=1,…,p) $H_0 : \beta_j = 0$

- statistica test $t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}^{\ 2}}} \sim t(n-p-1)$

- valutazione → il coefficiente è significativo (significativamente diverso da 0) se il corrispondente p-value è piccolo (ossia, rifiuto l'ipotesi di coefficiente nullo) → il regressore a cui il coefficiente è associato è rilevante per la spiegazione del fenomeno

Il modello di regressione lineare

La stima del modello

Root MSE	55693	**R-Square**	0.6207
Dependent Mean	32431	**Adj R-Sq**	0.6200
Coeff Var	171.72861		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	**1**	-15016	2324.86370	-6.46	<.0001
PAG_ORD	Pagato in contrassegno	**1**	1.19433	0.05485	21.78	<.0001
PAG_MES	Pagato con rate mensili	**1**	2.52341	0.10102	24.98	<.0001
TOT_ORD	Totale ordini	**1**	14881	683.88703	21.76	<.0001
LISTA	Numero di liste di appartenenza	**1**	603.36550	1110.84778	0.54	0.5871
SESSO	Sesso	**1**	3453.14705	1994.83468	1.73	0.0835
CEN	Residenza Centro	**1**	-6431.88493	2597.25872	-2.48	0.0133
SUD	Residenza Sud	**1**	-18390	2077.96317	-8.85	<.0001

Il modello di regressione lineare

La stima del modello

Interpretazione dei coefficienti

- impatto di X_j su Y posto che nel modello sono presenti altre variabili

- tasso di variazione di Y al variare di X_j

- come varia Y al variare di una unità di X_j se gli altri regressori non variano

Il modello di regressione lineare

La stima del modello

Segno del coefficiente

- indica la direzione dell'impatto del regressore a cui è associato
- segno atteso diverso da quello osservato può indicare interazione tra i regressori (multicollinearità)

Ordine di grandezza

- dipende dall'unità di misura della variabile indipendente Xj
- per valutarlo usare coefficienti standardizzati

Il modello di regressione lineare

La stima del modello

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	**1**	-14624	2205.46539	-6.63	<.0001	0	0
PAG_ORD	Pagato in contrassegno	**1**	1.15419	0.05482	21.05	<.0001	0.36897	2.96182
PAG_MES	Pagato con rate mensili	**1**	2.56876	0.09567	26.85	<.0001	0.27583	1.01781
TOT_ORD	Totale ordini	**1**	14434	674.26080	21.41	<.0001	0.37406	2.94467
LISTA	Numero di liste di appartenenza	**1**	872.66180	1052.55642	0.83	0.4071	0.00845	1.00196
SESSO	Sesso	**1**	3192.81846	1889.02931	1.69	0.0911	0.01726	1.00599
CEN	Residenza Centro	**1**	-6320.88855	2462.17857	-2.57	0.0103	-0.02792	1.14079
SUD	Residenza Sud	**1**	-17923	1971.41534	-9.09	<.0001	-0.10108	1.19214

Il modello di regressione lineare

La stima del modello

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	**1**	-14624	2205.46539	-6.63	<.0001	0	0
PAG_ORD	Pagato in contrassegno	**1**	1.15419	0.05482	21.05	<.0001	0.36897	2.96182
PAG_MES	Pagato con rate mensili	**1**	2.56876	0.09567	26.85	<.0001	0.27583	1.01781
TOT_ORD	Totale ordini	**1**	14434	674.26080	21.41	<.0001	0.37406	2.94467
LISTA	Numero di liste di appartenenza	**1**	872.66180	1052.55642	0.83	0.4071	0.00845	1.00196
SESSO	Sesso	**1**	3192.81846	1889.02931	1.69	0.0911	0.01726	1.00599
CEN	Residenza Centro	**1**	-6320.88855	2462.17857	-2.57	0.0103	-0.02792	1.14079
SUD	Residenza Sud	**1**	-17923	1971.41534	-9.09	<.0001	-0.10108	1.19214

Il modello di regressione lineare

La stima del modello

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	**1**	-14624	2205.46539	-6.63	<.0001	0	0
PAG_ORD	Pagato in contrassegno	**1**	1.15419	0.05482	21.05	<.0001	0.36897	2.96182
PAG_MES	Pagato con rate mensili	**1**	2.56876	0.09567	26.85	<.0001	0.27583	1.01781
TOT_ORD	Totale ordini	**1**	14434	674.26080	21.41	<.0001	0.37406	2.94467
LISTA	Numero di liste di appartenenza	**1**	872.66180	1052.55642	0.83	0.4071	0.00845	1.00196
SESSO	Sesso	**1**	3192.81846	1889.02931	1.69	0.0911	0.01726	1.00599
CEN	Residenza Centro	**1**	-6320.88855	2462.17857	-2.57	0.0103	-0.02792	1.14079
SUD	Residenza Sud	**1**	-17923	1971.41534	-9.09	<.0001	-0.10108	1.19214

Il modello di regressione lineare

La stima del modello

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	**1**	-14624	2205.46539	-6.63	<.0001	0	0
PAG_ORD	Pagato in contrassegno	**1**	1.15419	0.05482	21.05	<.0001	0.36897	2.96182
PAG_MES	Pagato con rate mensili	**1**	2.56876	0.09567	26.85	<.0001	0.27583	1.01781
TOT_ORD	Totale ordini	**1**	14434	674.26080	21.41	<.0001	0.37406	2.94467
LISTA	Numero di liste di appartenenza	**1**	872.66180	1052.55642	0.83	0.4071	0.00845	1.00196
SESSO	Sesso	**1**	3192.81846	1889.02931	1.69	0.0911	0.01726	1.00599
CEN	Residenza Centro	**1**	-6320.88855	2462.17857	-2.57	0.0103	-0.02792	1.14079
SUD	Residenza Sud	**1**	-17923	1971.41534	-9.09	<.0001	-0.10108	1.19214

Il modello di regressione lineare

1. Introduzione ai modelli di regressione – Case Study
2. Obiettivi
3. Le ipotesi del modello
4. La stima del modello
5. La valutazione del modello
   - Analisi della Multicollinearità
6. Commenti

Il modello di regressione lineare

La Multicollinearità

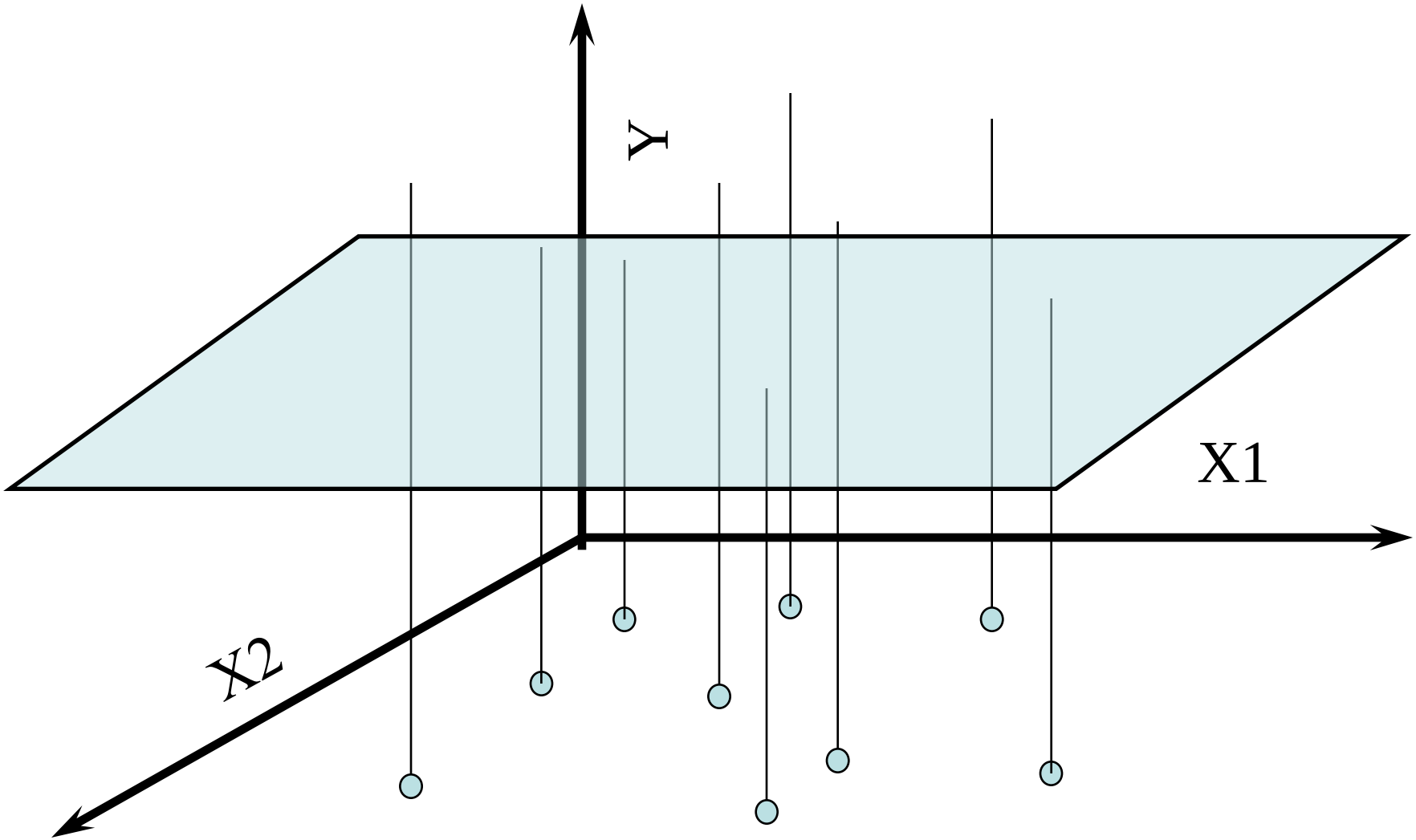
- $X_1,\ldots,X_p$ **non** sono vettori linearmente indipendenti
- forte correlazione tra i regressori (o alcuni di essi)

→ La varianza dello stimatore dei minimi quadrati tende ad esplodere

→ Problema di stabilità delle stime

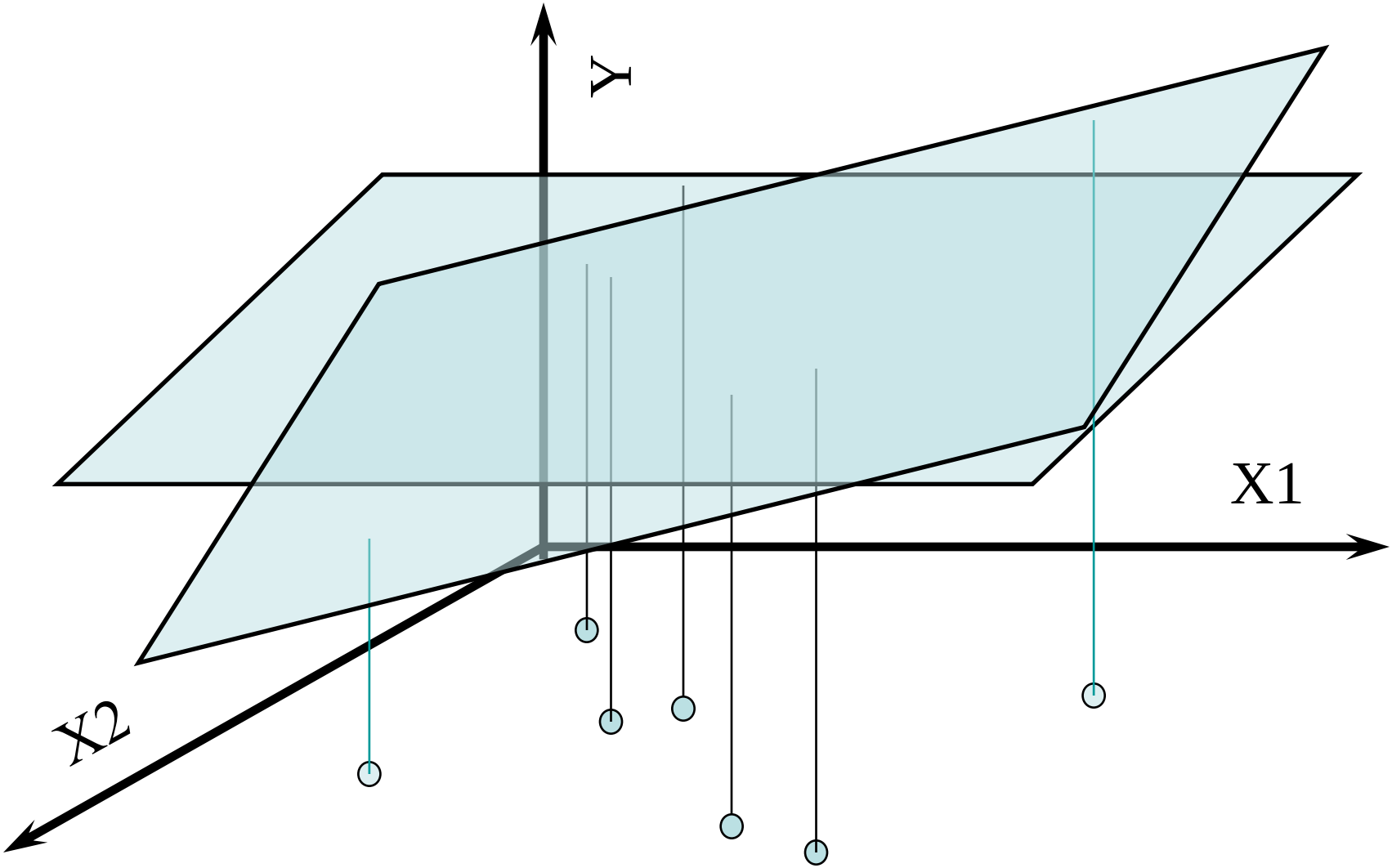
Il modello di regressione lineare

La Multicollinearità



Il modello di regressione lineare

La Multicollinearità



Il modello di regressione lineare

La Multicollinearità

R2	VIF
0.1	1.11
0.2	1.25
0.3	1.43
0.4	1.67
0.5	2.00
0.6	2.50
0.7	3.33
0.8	5.00
0.9	10.00
0.95	20.00
0.98	50.00
0.99	100.00

Per verificare la presenza di multicollinearità

• regressione lineare di X_j sui rimanenti p-1 regressori

- **R_j^2 -** misura la quota di varianza di X_j spiegata dai rimanenti p-1 regressori → **valori > 0.2 / 0.3 → presenza di multicollininearità**

- **Variance Inflation Index (VIF_j)**
  $VIF_j = 1 / (1 – R_j^2)$ misura il grado di relazione lineare tra X_j e i rimanenti p-1 regressori
  → valori > 1.2 / 1.3 → presenza di multicollininearità.

Il modello di regressione lineare

La Multicollinearità

Soluzioni

- trasformazione delle variabili correlate
- selezione di una variabile rappresentativa dal gruppo di variabili legate da relazione lineare e rimozione delle alre variabili correlate
- analisi delle componenti principali → trasformazione dei regressori in componenti non correlate (nella nuova regressione andranno incluse tutte le componenti principali)

Il modello di regressione lineare

La Multicollinearità

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	**1**	-14624	2205.46539	-6.63	<.0001	0	0
PAG_ORD	Pagato in contrassegno	**1**	1.15419	0.05482	21.05	<.0001	0.36897	2.96182
PAG_MES	Pagato con rate mensili	**1**	2.56876	0.09567	26.85	<.0001	0.27583	1.01781
TOT_ORD	Totale ordini	**1**	14434	674.26080	21.41	<.0001	0.37406	2.94467
LISTA	Numero di liste di appartenenza	**1**	872.66180	1052.55642	0.83	0.4071	0.00845	1.00196
SESSO	Sesso	**1**	3192.81846	1889.02931	1.69	0.0911	0.01726	1.00599
CEN	Residenza Centro	**1**	-6320.88855	2462.17857	-2.57	0.0103	-0.02792	1.14079
SUD	Residenza Sud	**1**	-17923	1971.41534	-9.09	<.0001	-0.10108	1.19214

Il modello di regressione lineare

La Multicollinearità

Root MSE	52693	**R-Square**	0.6204
Dependent Mean	30935	**Adj R-Sq**	0.6197
Coeff Var	170.33339		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	**1**	30935	869.91751	35.56	<.0001	0
Factor1		**1**	61162	870.03609	70.30	<.0001	1.00000
Factor2		**1**	-295.62943	870.03609	-0.34	0.7340	1.00000
Factor3		**1**	24154	870.03609	27.76	<.0001	1.00000
Factor4		**1**	3446.48124	870.03609	3.96	<.0001	1.00000
Factor5		**1**	861.78906	870.03609	0.99	0.3220	1.00000
Factor6		**1**	-13861	870.03609	-15.93	<.0001	1.00000
Factor7		**1**	73.57034	870.03609	0.08	0.9326	1.00000

Il modello di regressione lineare

1. Introduzione ai modelli di regressione – Case Study
2. Obiettivi
3. Le ipotesi del modello
4. La stima del modello
5. La valutazione del modello
   - La selezione dei regressori
6. Commenti

Il modello di regressione lineare

La selezione dei regressori

Poche variabili
- capacità previsiva ↓
- fit ↓
- parsimonia ↑
- interpretabilità ↑

Tante variabili
- capacità previsiva ↑
- fit ↑
- parsimonia ↓
- interpretabilità ↓

Criteri di selezione
- valutazioni soggettive
- confronto tra tutti i possibili modelli
- algoritmi di selezione automatica

Il modello di regressione lineare

La selezione dei regressori

Procedura di calcolo automatico che seleziona il sottoinsieme di variabili ottimo tra quelli possibili

• **forward selection** → inserisce nell'equazione una variabile per volta, basandosi sul contributo del regressore inserito alla spiegazione della variabilità di Y

• **backward selection** → rimuove dall'equazione una variabile per volta, basandosi sulla perdita di capacità esplicativa della variabilità di Y conseguente all'eliminazione del regressore

• **stepwise selection (**forward+backward selection) → ogni variabile può entrare/uscire dal modello

Il modello di regressione lineare

La selezione dei regressori

La **Stepwise Selection** è una procedura sequenziale che valuta l'ingresso/uscita dal modello dei singoli regressori (in base a indicatori legati all'R-quadro)

- **Step 0**→si considerano tutti i potenziali regressori
- **Step 1**→entra il primo regressore. Ossia, viene stimato un modello contenente un unico regressore tra quelli proposti (viene scelto il regressore che spiega meglio la variabilità della variabile dipendente)
- **Step 2**→si valutano tutti i possibili modelli contenenti il regressore individuato allo step 1 e uno dei rimanenti regressori, e si tiene il modello con il fit migliore (ossia entra il regressore che dà il contributo maggiore alla spiegazione della variabilità)

Il modello di regressione lineare

La selezione dei regressori

- **Step 3 e seguenti** →si valuta l'uscita di ognuno dei regressori presenti (in base alla minor perdita di capacità esplicativa del modello) e l'ingresso di un nuovo regressore (in base al maggior incremento nella capacità esplicativa del modello). Tra tutti i regressori rimanenti verrà scelto quello che dà il contributo maggiore alla spiegazione della variabilità della variabile dipendente

- **Ultimo step** → la procedura si arresta quando nessun regressore rimanente può essere inserito in base al livello di significatività scelto (slentry) e nessun regressore incluso può essere eliminato in base al livello di significatività scelto (slstay). In pratica quando non si riesce in alcun modo ad aumentare la capacità esplicativa del modello

Il modello di regressione lineare

La selezione dei regressori

Root MSE	52693	**R-Square**	0.6204
Dependent Mean	30935	**Adj R-Sq**	0.6197
Coeff Var	170.33339		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	**1**	30935	869.91751	35.56	<.0001	0
Factor1		**1**	61162	870.03609	70.30	<.0001	1.00000
Factor2		**1**	-295.62943	870.03609	-0.34	0.7340	1.00000
Factor3		**1**	24154	870.03609	27.76	<.0001	1.00000
Factor4		**1**	3446.48124	870.03609	3.96	<.0001	1.00000
Factor5		**1**	861.78906	870.03609	0.99	0.3220	1.00000
Factor6		**1**	-13861	870.03609	-15.93	<.0001	1.00000
Factor7		**1**	73.57034	870.03609	0.08	0.9326	1.00000

Il modello di regressione lineare

La selezione dei regressori

Root MSE	52679	**R-Square**	0.6203
Dependent Mean	30935	**Adj R-Sq**	0.6199
Coeff Var	170.28930		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	**1**	30935	869.69238	35.57	<.0001	0	0
Factor1		**1**	61162	869.81092	70.32	<.0001	0.71583	1.00000
Factor3		**1**	24154	869.81092	27.77	<.0001	0.28269	1.00000
Factor4		**1**	3446.48124	869.81092	3.96	<.0001	0.04034	1.00000
Factor6		**1**	-13861	869.81092	-15.94	<.0001	-0.16223	1.00000

Case study

Il prezzo e la spesa in attività promozionali sono due dei fattori che determinano le vendite di un prodotto.

Supponiamo che una grande catena di negozi alimentari operante su scala nazionale intenda introdurre una barretta energetica di basso prezzo.

Prima di introdurre il nuovo prodotto sul mercato si vuole stabilire l'effetto che il prezzo e le promozioni all'interno dei negozi possono avere sulle vendite.

Un campione di 34 negozi della catena viene selezionato per una ricerca di mercato. I negozi hanno tutti approssimativamente il medesimo volume di vendite mensili.

Case study

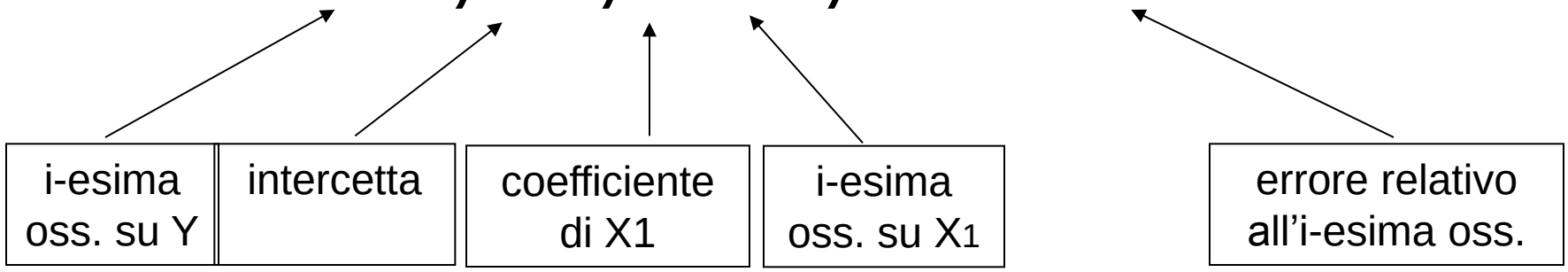
Si prendono in considerazioni due variabili indipendenti:

- il prezzo di una barretta (*X*1) e
- la spesa mensile per le attività promozionali (i cartelli pubblicitari, i tagliandi di sconto e i campioni gratuiti) (*X*2).

La variabile dipendente *Y* è il numero di barrette vendute in un mese.

Equazione di regressione lineare multipla

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$



Case study

Root MSE	638.06529	**R-Square**	0.7577
Dependent Mean	3098.6765	**Adj R-Sq**	0.7421
Coeff Var	20.59154		

Il coefficiente di determinazione è uguale a 0.7577 e, quindi, ci dice che il 75.77% della variabilità delle vendite è spiegato dal prezzo e dalle spese promozionali.

Considerando l' R^2 *corretto:* il 74.21% della variabilità delle vendite può essere spiegato dal modello proposto, tenuto conto delle numero di regressori e dell'ampiezza campionaria

Case study

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	39472731	19736365	48.48	<.0001
Error	31	12620947	407127		
Corrected Total	33	52093677			

Test F per la significatività congiunta dei coefficienti

H_0: $\beta_1 = \beta_2 = 0$

H_1 : Almeno un β_j <> 0

Se il livello di significatività scelto è 0.05, poiché il *p*-value è < 0.0001 e quindi < 0.05, possiamo rifiutare H_0 e quindi concludere che vi è una relazione lineare tra almeno una variabile esplicativa e la variabile dipendente (vendite)

Case study

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**	**Variance Inflation**
Intercept	Intercept	**1**	5837.5208	628.1502	9.29	<.0001	0	0
Prezzo	Prezzo	**1**	-53.21734	6.85222	-7.77	<.0001	-0.68982	1.00945
Promozione	Promozione	**1**	3.61306	0.68522	5.27	<.0001	0.46834	1.00945

Test t per la significatività dei singoli coefficienti

H_0: $\beta_i = 0$

H_1 : $\beta_i <> 0$

Se il livello di significatività scelto è 0.05, poiché il *p*-value è in entrami i casi < 0.0001 e quindi < 0.05, possiamo rifiutare H_0 e quindi concludere che entrambe le variabili sono significative alla spiegazione del fenomeno

Case study

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**	**Variance Inflation**
Intercept	Intercept	**1**	5837.5208	628.1502	9.29	<.0001	0	0
Prezzo	Prezzo	**1**	-53.21734	6.85222	-7.77	<.0001	-0.68982	1.00945
Promozione	Promozione	**1**	3.61306	0.68522	5.27	<.0001	0.46834	1.00945

$$VIF_j = \frac{1}{1 - R_j^2}$$

I valori del **Variance Inflation Index** minori di 2 garantiscono l'assenza di multicollinearità.

Case study

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**	**Variance Inflation**
Intercept	Intercept	**1**	5837.5208	628.1502	9.29	<.0001	0	0
Prezzo	Prezzo	**1**	-53.21734	6.85222	-7.77	<.0001	-0.68982	1.00945
Promozione	Promozione	**1**	3.61306	0.68522	5.27	<.0001	0.46834	1.00945

I coefficienti di regressione sono legati all'unità di misura delle variabili. Ciò significa che la grandezza di un particolare coefficiente non è un buon indicatore della sua importanza.

I **coefficienti standardizzati** sono utili per valutare l'importanza relativa dei regressori. Possiamo ordinare i regressori in base all'importanza che hanno nello spiegare la variabile dipendente.

Il regressore con valore assoluto del coefficiente standardizzato più alto è il più importante.

Nell'esempio il prezzo è il regressore più importante (|-0.69|) e poi la spesa mensile per le attività promozionali (|0.47|)

Case study

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**	**Variance Inflation**
Intercept	Intercept	**1**	5837.5208	628.1502	9.29	<.0001	0	0
Prezzo	Prezzo	**1**	-53.21734	6.85222	-7.77	<.0001	-0.68982	1.00945
Promozione	Promozione	**1**	3.61306	0.68522	5.27	<.0001	0.46834	1.00945

Una volta stimati i valori dei parametri della regressione la previsione viene calcolata semplicemente applicando la relazione lineare:

$$\hat{Y}_i = 5837.52 - 53.2173 X_{1i} + 3.6131 X_{2i}$$

I coefficienti in un modello di regressione multipla misurano la variazione della variabile risposta *Y* in corrispondenza della variazione di una delle variabili esplicative, quando si tengono costanti le altre.

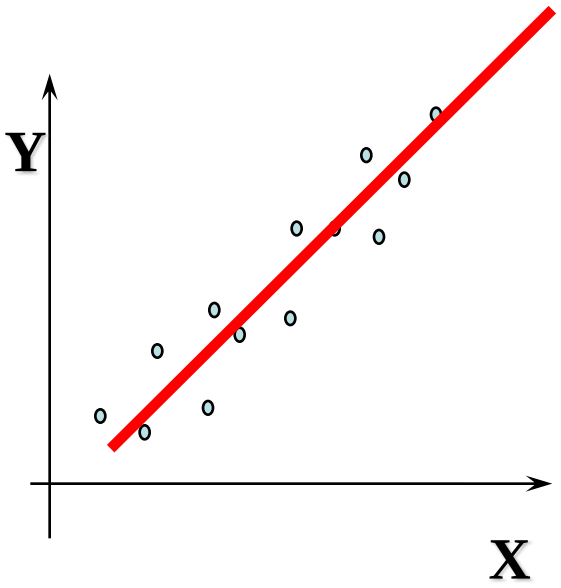
Il modello di regressione lineare

1. Introduzione ai modelli di regressione – Case Study
2. Obiettivi
3. Le ipotesi del modello
4. La stima del modello
5. La valutazione del modello
6. Commenti

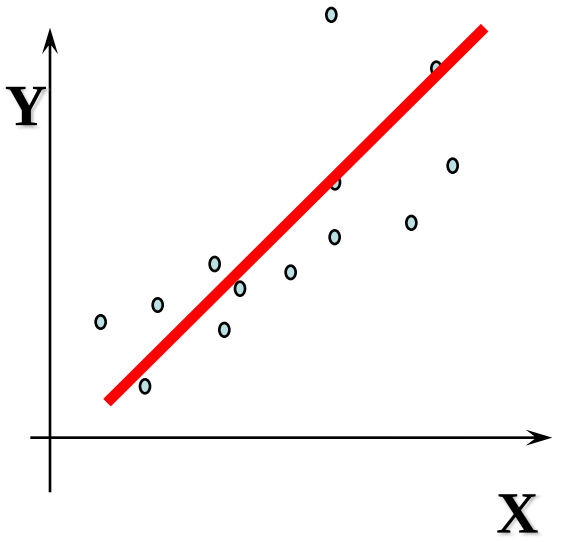
Il modello di regressione lineare

La stima del modello

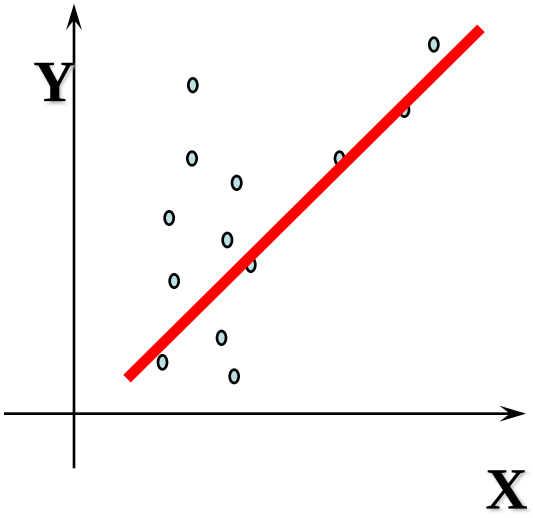
Indicatori di bontà del Modello



R-SQUARE=0.7
F con p-value piccolo



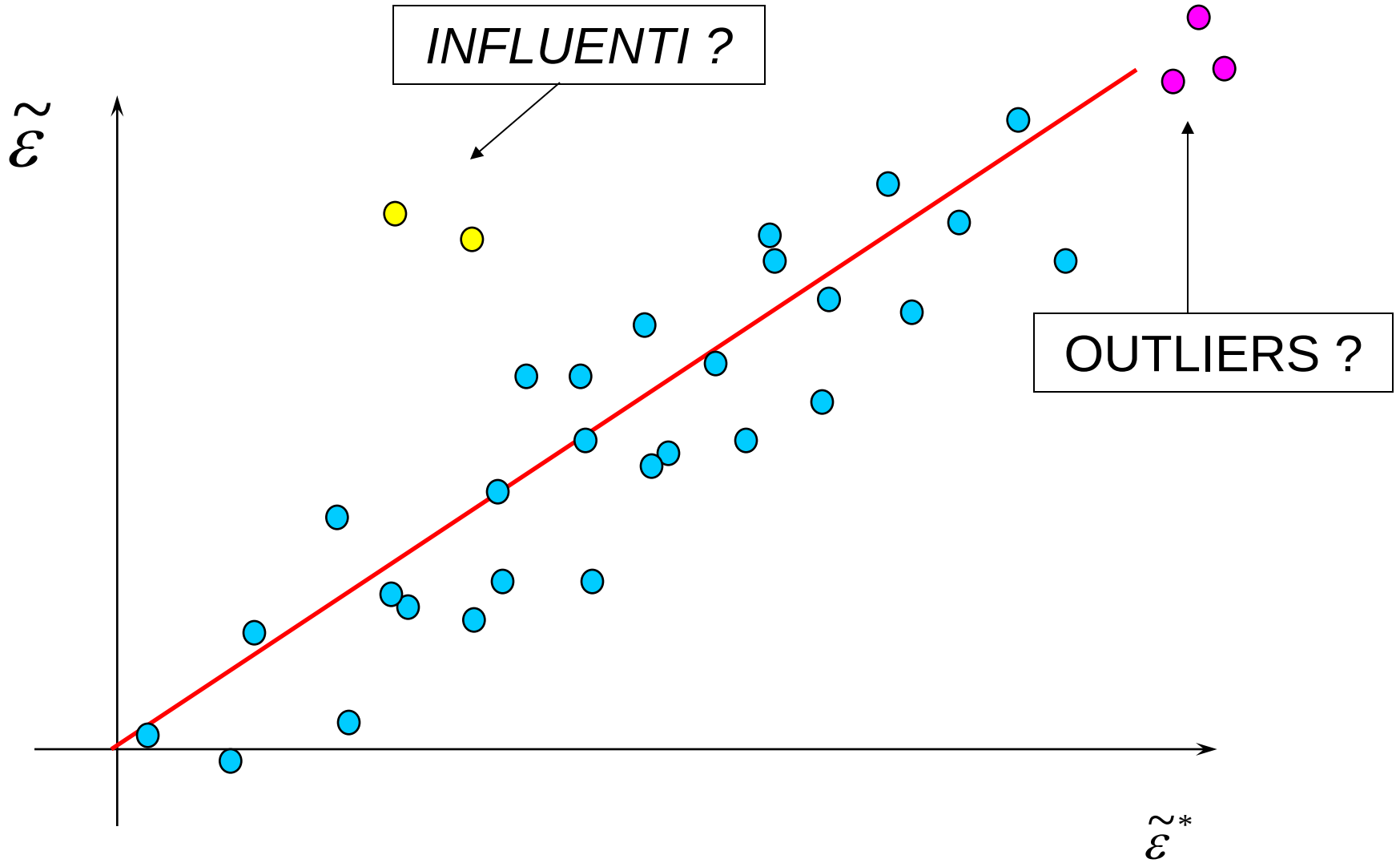
R-SQUARE=0.7
F con p-value piccolo



R-SQUARE=0.7
F con p-value piccolo

Il modello di regressione lineare

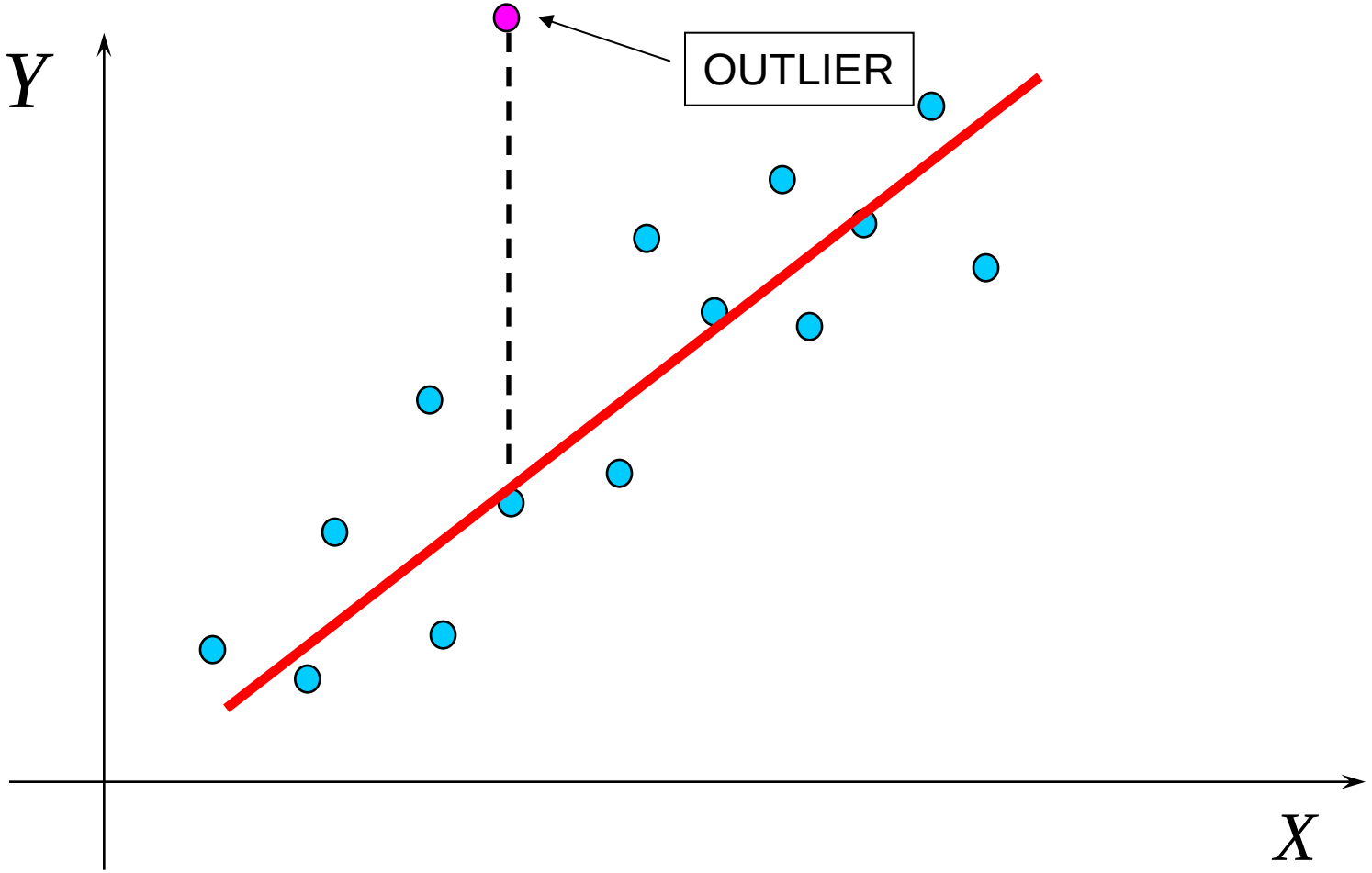
L'analisi di Influenza



Il modello di regressione lineare

L'analisi di Influenza

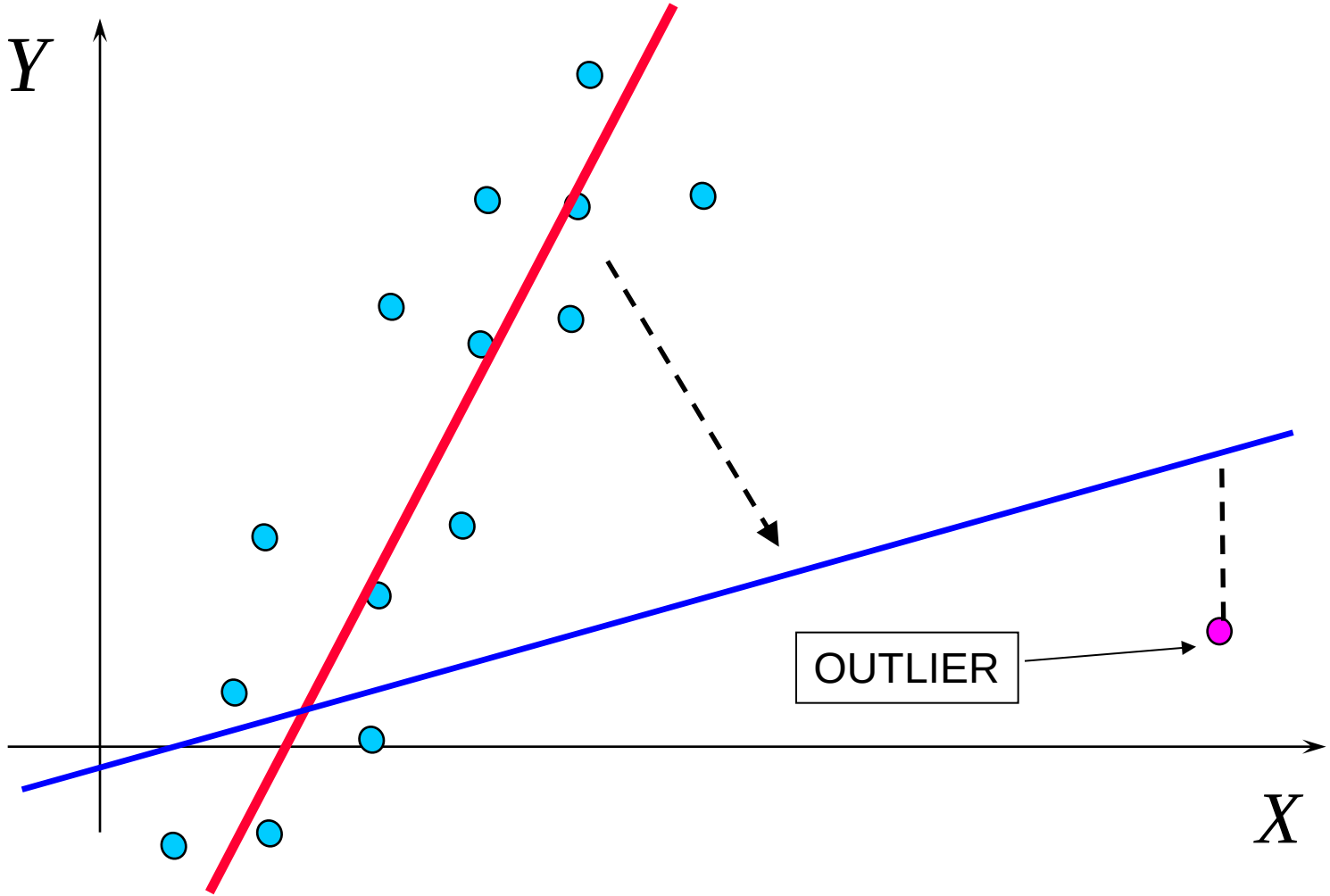
Osservazione anomala rispetto alla variabilità di Y → non attira a sé il modello in maniera significativa



Il modello di regressione lineare

L'analisi di Influenza

Osservazione anomala rispetto alla variabilità di Y→ attira a sé il modello in maniera significativa



Il modello di regressione lineare

L’analisi di Influenza

Valutazione dell’impatto delle singole osservazioni

• osservazioni outlier che creano distorsione nella stima del modello

- plot dei residui
- plot X/Y

• osservazioni influenti che contribuiscono in modo “sproporzionato” alla stima del modello

- plot dei residui
- statistiche di influenza

Il modello di regressione lineare

Statistiche di Influenza

Leverage H: i-esimo elemento della diagonale della matrice di proiezione. misura quanto un'osservazione è lontana dal centro dei dati (ma tende a segnalare troppe oss influenti e tratta tutti i regressori nello stesso modo)

→ oss influente se lev H>2*(p+1)/n

$$[diag(H)]_i = [diag(X(X'X)^{-1}X')]_i$$

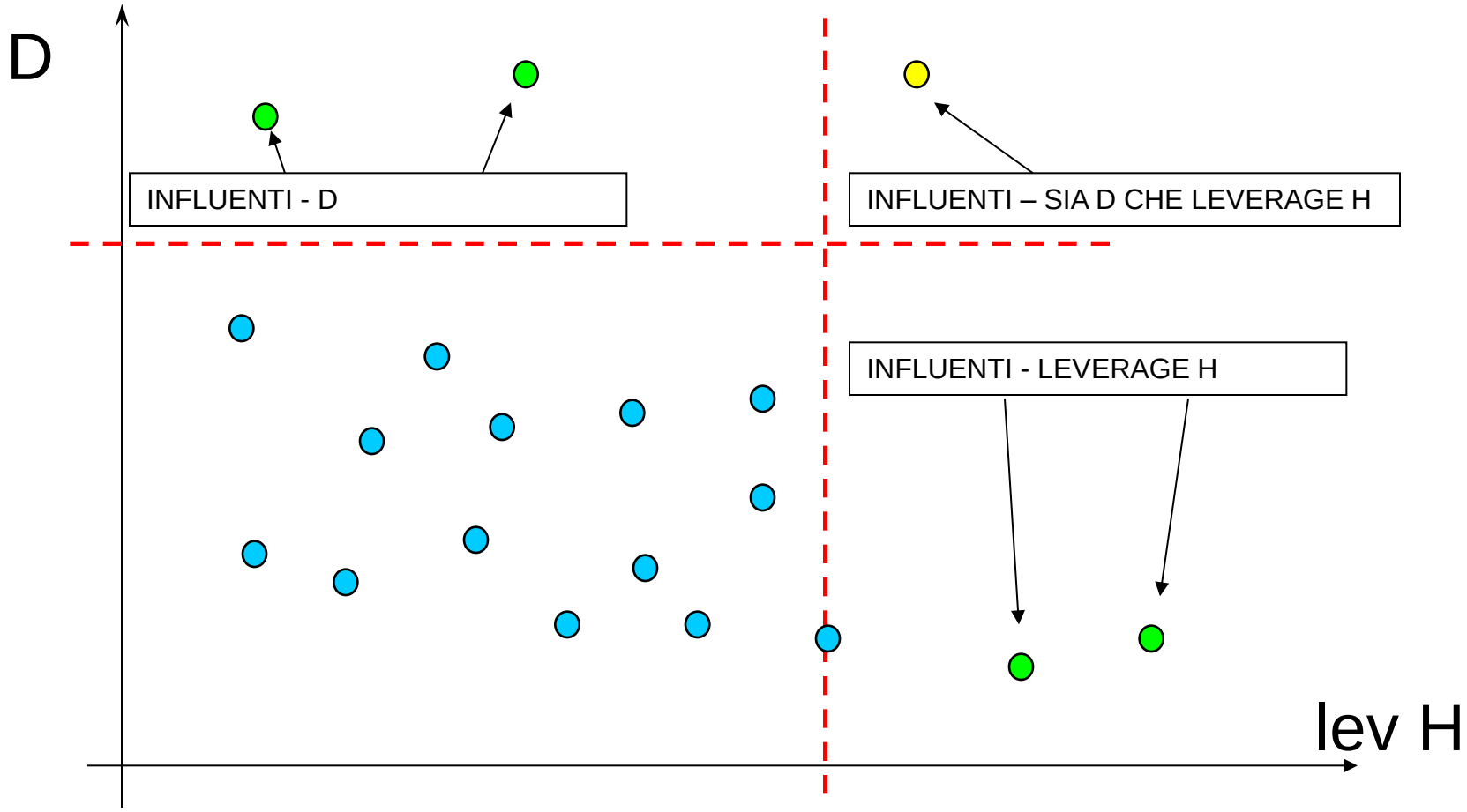
Distanza di Cook: misura la variazione simultanea dei coefficienti quando un'osservazione viene rimossa

→ oss influente se D>1

Il modello di regressione lineare

Statistiche di Influenza

Plot delle statistiche di influenza → attenzione alle osservazioni nel quadrante in alto a destra



Il modello di regressione lineare

Statistiche di Influenza

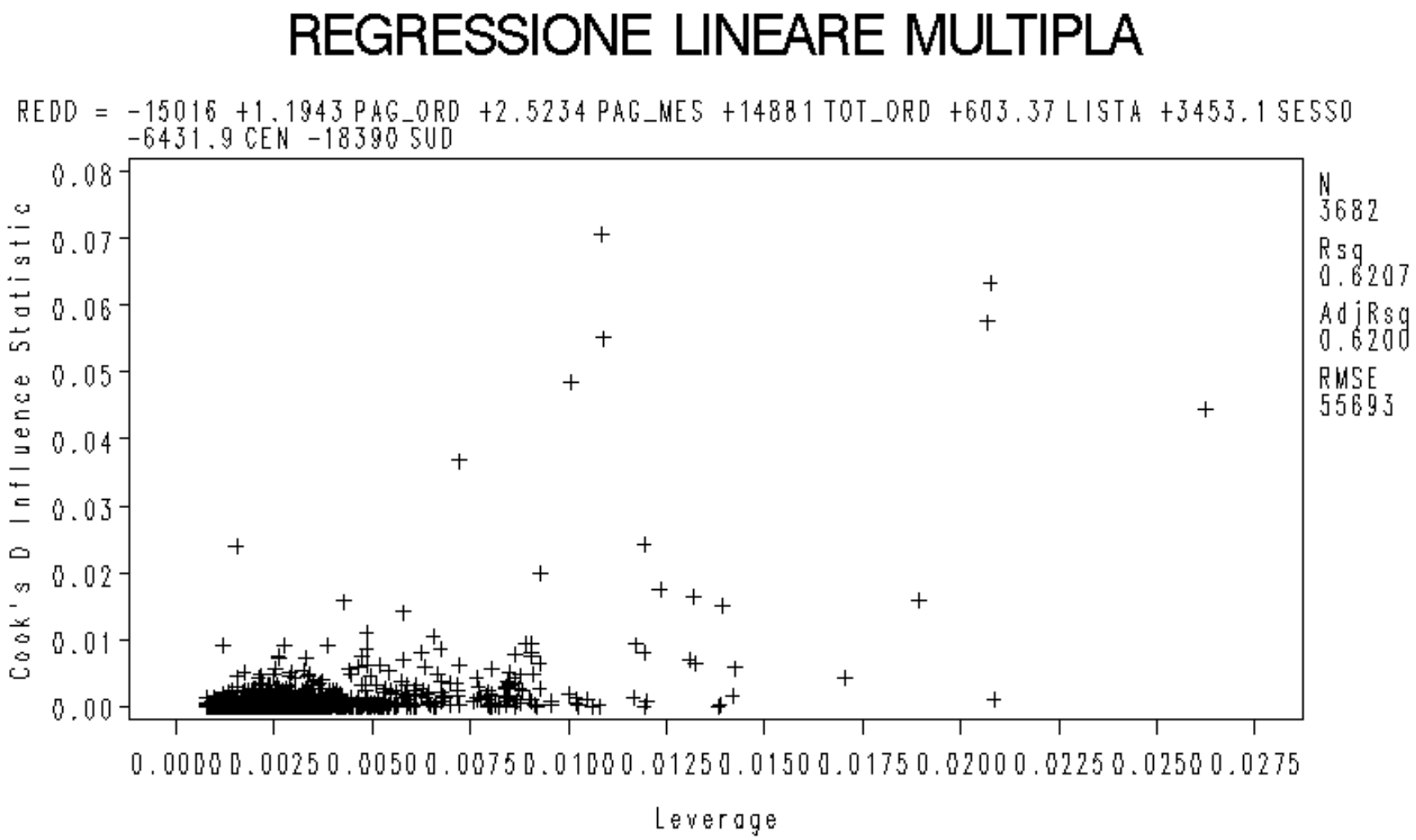
Root MSE	55693	**R-Square**	0.6207
Dependent Mean	32431	**Adj R-Sq**	0.6200
Coeff Var	171.72861		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	**1**	-15016	2324.86370	-6.46	<.0001
PAG_ORD	Pagato in contrassegno	**1**	1.19433	0.05485	21.78	<.0001
PAG_MES	Pagato con rate mensili	**1**	2.52341	0.10102	24.98	<.0001
TOT_ORD	Totale ordini	**1**	14881	683.88703	21.76	<.0001
LISTA	Numero di liste di appartenenza	**1**	603.36550	1110.84778	0.54	0.5871
SESSO	Sesso	**1**	3453.14705	1994.83468	1.73	0.0835
CEN	Residenza Centro	**1**	-6431.88493	2597.25872	-2.48	0.0133
SUD	Residenza Sud	**1**	-18390	2077.96317	-8.85	<.0001

Il modello di regressione lineare

Statistiche di Influenza



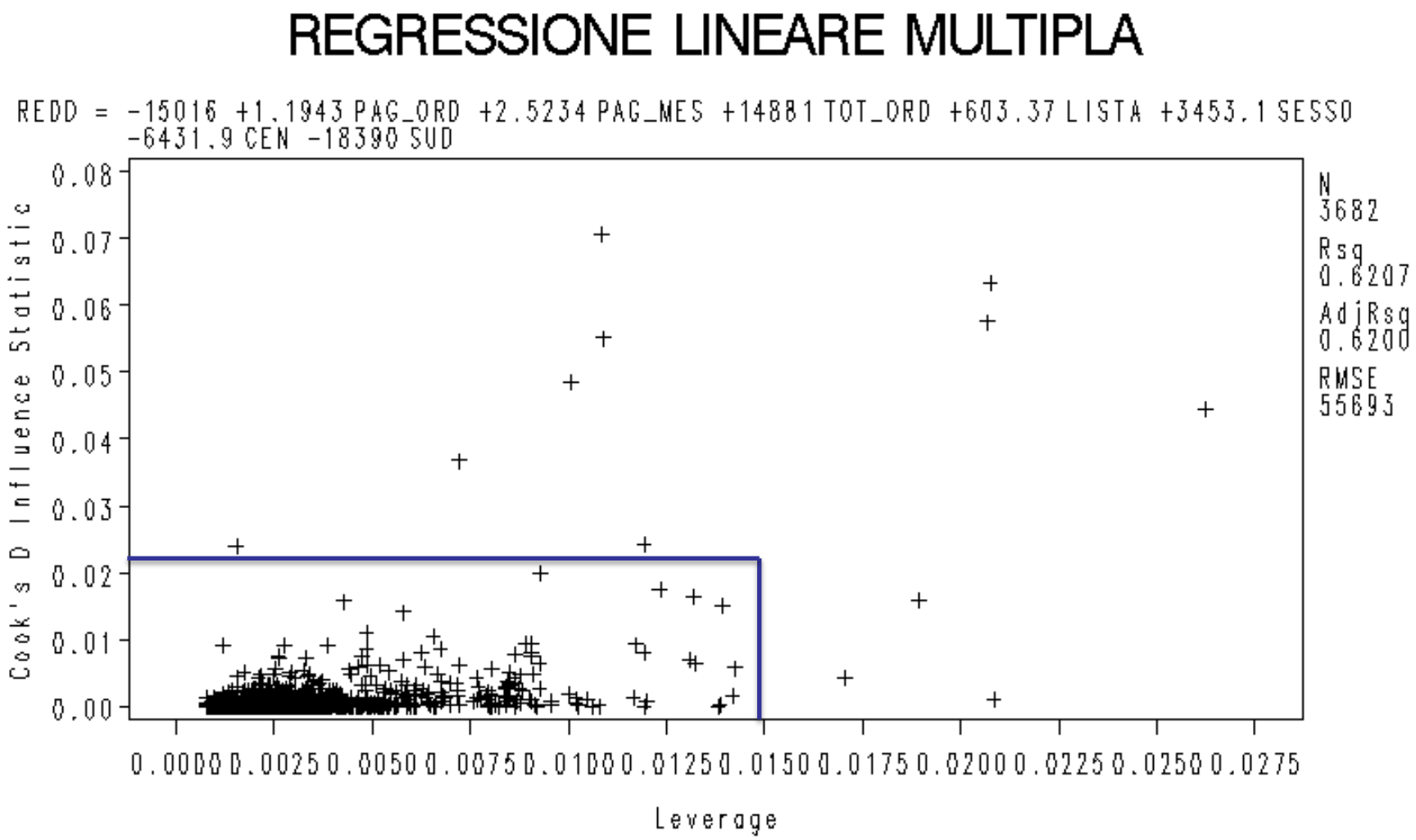
Il modello di regressione lineare

Statistiche di Influenza

```
DATA REGRESS1 (DROP = COOK H REDD_PRE RES_STUD);
SET  RESID_0;

WHERE COOK < 0.023 & H < 0.015;

PROC REG DATA=REGRESS1;
MODEL      REDD=PAG_ORD PAG_MES TOT_ORD LISTA
           SESSO CEN SUD ;

PAINT RSTUDENT.> 2 / SYMBOL='O';
PAINT RSTUDENT.<-2 / SYMBOL='O';

PLOT RSTUDENT.*P.;
PLOT       P.*REDD;
PLOT    COOKD.*H.;

RUN;
```

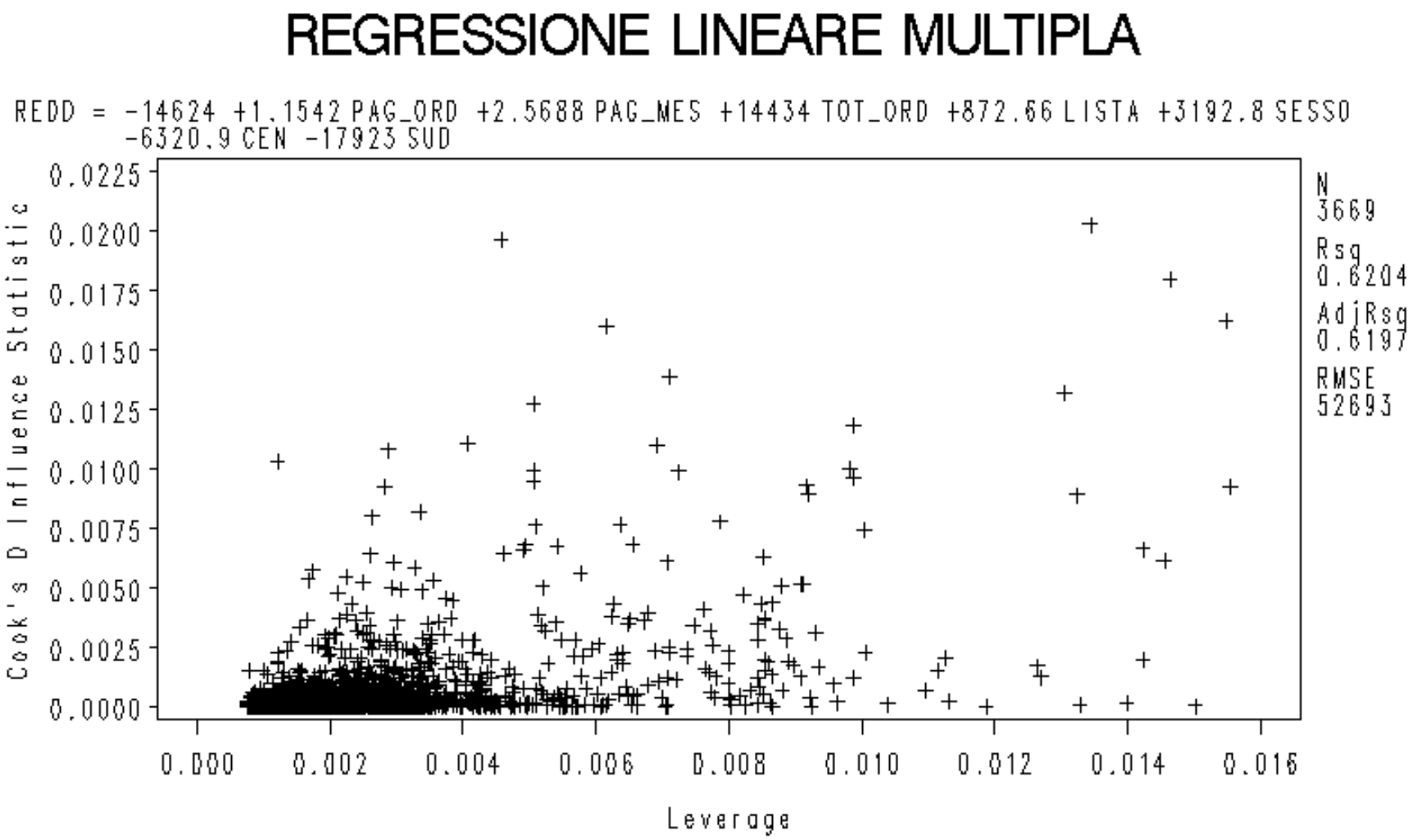
Il modello di regressione lineare

Statistiche di Influenza



Il modello di regressione lineare

Statistiche di Influenza



Il modello di regressione lineare

Statistiche di Influenza

Root MSE	52693	**R-Square**	0.6204
Dependent Mean	30935	**Adj R-Sq**	0.6197
Coeff Var	170.33339		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	**1**	-14624	2205.46539	-6.63	<.0001
PAG_ORD	Pagato in contrassegno	**1**	1.15419	0.05482	21.05	<.0001
PAG_MES	Pagato con rate mensili	**1**	2.56876	0.09567	26.85	<.0001
TOT_ORD	Totale ordini	**1**	14434	674.26080	21.41	<.0001
LISTA	Numero di liste di appartenenza	**1**	872.66180	1052.55642	0.83	0.4071
SESSO	Sesso	**1**	3192.81846	1889.02931	1.69	0.0911
CEN	Residenza Centro	**1**	-6320.88855	2462.17857	-2.57	0.0103
SUD	Residenza Sud	**1**	-17923	1971.41534	-9.09	<.0001

Il modello di regressione lineare

La Valutazione del modello

Si vuole verificare

- bontà delle stime
- adattamento del modello ai dati
- impatto delle singole osservazioni
- impatto dei regressori

Strumenti

- test statistici
- indicatori di performance
- analisi dei residui
- analisi degli outliers
- analisi di influenza
- valutazione dei coefficienti e correlazioni parziali