Russell 3000 Index - Analisi di Regressione

La stima dell'EPS futuro è un'informazione fondamentale per un investitore: permette sia di indurre il potenziale ritorno sull'investimento ottenibile investendo in una specifica società, sia di comparare la redditività tra aziende diverse.

La regressione è una tecnica di analisi statistica che permette di indagare relazioni di dipendenza tra variabili e consente di studiare i rapporti causa-effetto: attraverso la regressione è possibile analizzare la dinamica dei profitti delle aziende e stimare l'EPS futuro di un'impresa, questo significa poter decidere se far parte del suo azionariato oppure no.

Le informazioni disponibili su 2482 [1] società presenti nell'indice Russell 3000 comprendono le rispettive categorie e sottocategorie industriali di appartenenza, oltre ai valori di alcuni indicatori economici e finanziari. Il dataset presenta una notevole variabilità in ognuno degli indicatori, con differenze marcate in base al settore industriale. Ad esempio, le società appartenenti al settore "Consumer Discretionary", in media, hanno elevati livelli di ROE e di Vendite Nette, mentre il settore "Healthcare" presenta valori mediamente tra i più bassi per tutti gli indicatori. Diverso invece per le società classificate in "Basic Materials" e "Industrials", le quali hanno ROE, ROA e ROIC generalmente molti alti.

Abbiamo sviluppato un modello di regressione ad hoc con l'obiettivo principale di comprendere la dinamica dei profitti aziendali, minimizzando sia l'errore di generalizzazione che il numero di regressori, eliminando quelli meno informativi.

Partendo dalle variabili disponibili, e considerando anche le loro combinazioni, abbiamo implementato l'algoritmo LASSO per la selezione delle variabili più significative. Alla luce delle suddette analisi, il modello che meglio rappresenta il fenomeno è [2]:

$$EPS = 0.02 + 0.19EV + 0.22ROA - 0{,}21(EV * ROE) + 0.41(NS * ROA) + \beta_i * Industry$$

Il modello contiene informazioni quantitative sugli indicatori maggiormente influenti rispetto alle stime dell'EPS futuro (come l'Enterprise Value, il Return On Asset, e le interazioni presenti tra EV - ROE e NS – ROA), ma anche le informazioni qualitative sulle categorie industriali.

Nell'equazione del modello è presente l'intercetta, all'interno della quale troviamo anche informazioni riguardo il settore "Basic Materials" [3].

Dai coefficienti del modello possono essere estratte le seguenti informazioni:

Industry	Betas
Consumer Discretionary	0.12
Consumer Staples	-0.15
Energy	0.06
Financials	0.08
Healthcare	-0.21
Industrials	0.01
Real Estate	-0.27
Technology	-0.14
Telecommunications	-0.18
Utilities	-0.17

- Enterprise Value (EV) e Return On Assets (ROA)
  Entrambi i coefficienti (EV: + 0.19; ROA: + 0.22) segnalano un impatto positivo delle variabili rispetto all'EPS stimato. Nello specifico, il modello individua una relazione lievemente più forte tra ROA ed EPS. Dunque, in fase di valutazione dell'EPS futuro stimato per prendere decisioni di business, sarebbe opportuno dare maggiore importanza alle società con ROA elevato (e possibilmente in crescita rispetto al passato), in ogni caso senza perdere l'attenzione sui suoi valori di EV.

In aggiunta alle variabili fornite, sono state considerate combinazioni lineari tra le stesse, delle quali si sono rivelate maggiormente significative le seguenti:

- Enterprise Value • Return On Equity
  Il valore dell'impresa, moltiplicato per la redditività del capitale investito, ha un'influenza negativa sull'EPS: ciò significa che all'aumentare di questo prodotto di una unità, l'EPS diminuisce di 0.21, mantenendo costanti gli effetti delle altre variabili.

- Net Sales • Return On Asset
  Le vendite nette dell'impresa, moltiplicate per la redditività della stessa in relazione alle risorse utilizzate per svolgere la propria attività economica ha un'influenza positiva sull'EPS. Questo significa che all'aumentare di questo prodotto l'EPS aumenta di 0,41.

Al fine di utilizzare il modello per scopi inferenziali, cioè prevedere in modo affidabile i valori futuri della variabile target, è necessario che lo stimatore OLS del fenomeno analizzato sia consistente. In questo caso specifico, i coefficienti stimati ci permettono di spiegare la variabile dipendente ma non di fare previsioni in modo efficiente e robusto [4].

Il modello in azione:

Il potenziale del modello può essere chiarito meglio presentando un esempio concreto: una società appartenente al settore "Tecnologia", con un EV di 6,59, Vendite nette a 9718, ROE di -4,21 e ROA di -0,12, avrà un EPS futuro stimato dal modello pari a 2,477 (mentre il valore effettivamente registrato di EPS è stato di 2,54).

Sviluppi Futuri:

Considerando i limiti di cui sopra, sono necessari ulteriori approfondimenti per poter sviluppare un'analisi robusta: la distorsione da variabili omesse può essere mitigata includendo nuove variabili che ben si prestano a descrivere il fenomeno come, per esempio, il numero di azioni circolanti. Includendo più istanze storiche per ciascuna variabile si potrebbe ridurre la distorsione da campionamento, andando ad indagare il fenomeno nel suo complesso piuttosto che limitandosi al semplice anno in esame.

Damian Agachi Menna
Annalisa Basta
Simone De Bonis
Chiara Mercati
Daniele Torregiani

Appendice

[1] Valori duplicati e anomali:

Durante l'analisi sono emersi 11 valori duplicati, per la stessa società erano presenti due righe con valori uguali per tutti gli attributi, con differenze solo per il campo "Net_Sales", e per l'analisi è stato selezionato il valore più realistico. Inoltre, è stato osservato che per determinati attributi sono presenti valori molto grandi e lontani dalla realtà, abbiamo quindi sostituito il valore anomalo con una media della sottocategoria dell'industria riferita a quel determinato attributo.

Ridondanza:

Durante l'analisi descrittiva, attraverso un correlogramma, abbiamo notato che alcune variabili erano fortemente correlate: abbiamo comunque deciso di prenderle in considerazione per l'analisi, dal momento che, come ci aspettavamo, le variabili ridondanti sarebbero state rimosse durante la fase di feature selection con la regressione LASSO.

[2] Coefficienti regressione:

La costruzione del modello ha richiesto diversi passaggi:

- Le variabili sono state standardizzate, dopodiché sono stati calcolati i prodotti incrociati, i quadrati e i cubi delle stesse;
- È stato utilizzato uno stimatore Lasso per identificare le variabili più importanti, l'iperparametro λ è stato individuato attraverso una 10-fold cross-validation con l'algoritmo Cyclical Coordinate Descent (CCD);
- Sulle variabili individuate dalla regressione LASSO è stato utilizzato un OLS.

[3] Multicollinearità:

Per evitare il problema della multicollinearità, nel modello sono state inserite n-1 variabili dummy (dove n è il numero di settori presenti), perciò il β relativo a Basic Materials è contenuto nell'intercetta, e tutte le altre variabili sono espresse in funzione della distanza da quest'ultima. Un coefficiente negativo associato ad uno degli altri settori individua una peggiore performance generale rispetto a Basic Materials; viceversa, in presenza di un coefficiente positivo.

[4] Consistenza dei coefficienti:

Per far sì che uno stimatore sia consistente devono essere verificate le condizioni di indipendenza lineare tra i regressori, una distribuzione omoschedastica dei residui e le variabili esplicative devono essere incorrelate con gli errori del modello.Data la complessità e la variabilità del fenomeno oggetto di studio, basandosi esclusivamente sulle covariate disponibili, si segnala un problema di potenziale mispecificazione (il RESET test rifiuta l'ipotesi nulla per ogni possibile combinazione di variabili, anche non lineare fino al terzo ordine) e, potenzialmente, di variabili omesse. Questo condurrebbe alla non consistenza dello stimatore OLS. Ciò significa che le stime del modello sono comunque indicative allo scopo di spiegare l'influenza che hanno le covariate sulla dinamica dell'EPS futuro, ma significa anche che i valori reali potrebbero scostarsi da quelli stimati. Per questo motivo abbiamo evitato di eseguire test di inferenza statistica sui parametri stimati.