**Progetto di analisi dati: Analisi del NASDAQ Composite**

**Corso: modelli e metodi dell’Inferenza Statistica**

**Chiara**

Abbiamo voluto analizzare il prezzo di apertura dell’indice NASDAQ a partire dai prezzi di apertura di alcuni dei titoli quotati nella borsa americana elettronica.

Per farlo abbiamo utilizzato un dataset contenente 253 osservazioni che rappresentano il prezzo di apertura delle azioni dal 20 maggio 2020 al 20 maggio 2021 per 12 colonne, sulla prima abbiamo le date, dalla seconda alla dodicesima abbiamo l’indice e i titoli di mercato. Analizziamo l'indice di mercato Nasdaq in correlazione con i suoi titoli Apple, Amazon, Microsoft, Starbucks, Google, Netflix, Tesla, PayPal, JD, Peloton.

Nella nostra analisi siamo partiti da una breve descrizione di ogni colonna del nostro dataset per visualizzare sinteticamente i prezzi di apertura di ogni titolo scelto.

Successivamente siamo andati a visualizzare graficamente come le variabili scelte sono relazionate tra loro. Attraverso il grafico ggpairs abbiamo analizzato la correlazione a coppie fra predittori e risposta. Già da ora si può notare che risulta esserci una collinearità positiva tra le variabili, che risulta anche molto marcata in alcuni casi. In seguito, osserveremo meglio come queste variabili sono correlate.

Chiara

Come prima analisi abbiamo studiato il modello di regressione lineare multipla scegliendo come variabile dipendente Y il prezzo dell’indice di mercato NASDAQ, mentre le variabili indipendenti Xi sono i prezzi di tutti i vari titoli citati sopra. Da questa prima regressione che comprende tutte le covariate abbiamo ottenuto un modello che risulta molto buono con un R2adj del 99% e un pvalue della statistica test molto basso, quindi molto significativo. Le covariate Apple, Starbucks, Google, Tesla, PayPal hanno un pvalue della statistica test molto basso e risultano quindi molto significative. Leggermente con un pvalue più alto sono Microsoft e Netflix ma risultano comunque molto significative. Mentre Amazon con un pvalue del 12% non risulta significativo, così come JD e Peloton con un pvalue rispettivamente del 13% e 43%.

A questo punto siamo andati a controllare le ipotesi del modello. Come prima cosa l’omoschedasticità dei residui risulta confermata poiché dal grafico non notiamo particolari andamenti e risulta una nuvola abbastanza omogenea. Possiamo quindi accettare l’ipotesi di omoschedasticità. Ora verifichiamo la normalità dei residui attraverso il qqplot e lo shapiro test. Entrambi i metodi ci permettono di confermare la normalità dei residui, questo perché nel qqplot otteniamo un andamento rettilineo con solo qualche valore che si discosta leggermente lungo le code. Nello Shapiro test invece otteniamo un pvalue alto del 39% che non ci permette di rifiutare l’ipotesi nulla di normalità.

Successivamente siamo andati a verificare l’ipotesi di non collinearità per evitare che due covariate diverse diano la stessa informazione e abbiamo calcolato il Variance Inflation Factor. Otteniamo dei valori molto alti, ciò significa che c’è forte dipendenza tra le covariate, come avevamo già osservato nel grafico iniziale. In particolare, Google mostra l'indice vif più alto di tutte. Dobbiamo quindi selezionare le covariate necessarie escludendo quelle superflue che andranno soltanto ad aumentare la varianza degli stimatori. Partiamo dal modello con tutte le covariate e utilizziamo il metodo automatico di selezione backward che toglie una covariata alla volta minimizzando gli indici AIC e BIC. Minimizzando l’AIC il modello ottenuto con la backward selection rimuove soltanto Peloton e non si risolve il problema di collinearità. Se utilizziamo lo stesso metodo di backward selection minimizzando invece il BIC vengono rimossi Peloton, JD e Amazon. Tuttavia, se facciamo il vif, per calcolare la collinearità otteniamo dei valori ancora elevati. Con il metodo “both” otteniamo lo stesso modello ottenuto con il metodo backward.

Niccolò

Procediamo quindi manualmente rimuovendo le covariate una alla volta dal modello iniziale completo. Partiamo inizialmente da quelle non significative, quindi Peloton, poi Amazon, poi JD. A questo punto poiché ci troviamo ancora in presenza di forte dipendenza tra le covariate continuiamo ad eliminarle togliendo quelle che hanno valori di indice vif maggiore. Rimuoviamo quindi Google, e successivamente Netflix che nel frattempo è diventata non significativa. Rimuoviamo poi anche PayPal che risulta tra le rimaste quella con il vif maggiore. Otteniamo a questo punto un modello le cui covariate rimanenti sono Apple, Microsoft, Starbucks, Tesla. Queste risultano tutte molto significative e dai valori di vif divenuti ora accettabili e non troppo elevati abbiamo un modello che finalmente non ha problemi di collinearità. Verifichiamo anche le altre ipotesi per questo modello, ovvero normalità ed omoschedasticità dei residui e notiamo che risultano confermate. Infatti, otteniamo uno shapiro test per i residui con pvalue del 15% e una nuvola abbastanza omogenea. Abbiamo fatto questa scelta di covariate, oltre ai motivi specificati sopra ma anche perché comprendono mercati diversi fra loro, ambito tech, ambito di ristorazione e di mobilità.

Abbiamo fatto un tentativo, dato che il qqplot non veniva troppo lineare nelle code, di trasformazione boxcox, trasformando quindi la Y variabile risposta e abbiamo ottenuto un lambda ottimale pari a 1,272727. Studiando i residui del nuovo modello trasformato non abbiamo ottenuto risultati molto migliori, con un pvalue dello shapiro test per i residui del 19% ma un qqplot ancora non troppo lineare nelle code. Per questo quindi ritorniamo a considerare il modello di prima della trasformazione boxcox dato che risulta essere un modello valido e molto buono.

Marco

Successivamente siamo andati a studiare la diagnostica dei punti influenti. Come prima cosa siamo andati a cercare i punti leva. Abbiamo quindi trovato la matrice H ed estratto i leverage, verificando poi che la somma faccia 5 ovvero il numero di covariate + 1. Abbiamo successivamente trovato quali punti eccedono il valore soglia 2p/n=0.03952569, riportandoli ed evidenziandoli in un grafico. La somma di tali punti leva risulta 0.3475209 su 5, quindi sono poco influenti. Provando a fittare il modello senza i punti di leva otteniamo un modello quasi invariato, ancora buono con un R2adj leggermente inferiore ma ancora molto elevato. Confrontando le stime dei coefficienti dei due modelli non abbiamo cambiamenti significativi dei beta hat, solo Microsoft del 1,3% e Starbucks 1,0%. Abbiamo poi analizzato i residui standardizzati cercando anche qui quali fossero i punti influenti cioè quei residui tali che | rist | > 2. Abbiamo confrontato graficamente i punti influenti ottenuti dai residui standardizzati con i punti leva trovati precedentemente notando che nei due casi troviamo punti influenti differenti. Abbiamo poi utilizzato lo stesso procedimento con i residui studentizzati trovando gli stessi punti influenti appena trovati con un unico punto influente in più. E infine abbiamo analizzato i punti influenti anche dal punto di vista della distanza di Cook e controllato come cambia il modello senza i punti influenti identificati. Osserviamo che l’ R2adj migliora leggermente e valutando le variazioni percentuali dei beta hat tra i due modelli si ha una leggera variazione con Apple che cambia dell’11% gli altri meno, ma si ha una variazione maggiore rispetto a quella ottenuta eliminando i levarage. Infine in un unico grafico abbiamo riportato sinteticamente i vari metodi, avendo i hii lungo le x, i res studentizzati lungo le y con i valori delle soglie tratteggiati e i cerchi che sono proporzionali alla distanza di Cook. Notando in quali date si hanno valori che influenzano particolarmente il modello.

Come ultima analisi abbiamo osservato come dando in input dei valori per i prezzi dell’ultimo modello scelto, quello con le 4 covariate, è possibile ottenere un intervallo di previsione e uno di confidenza. In particolare abbiamo dato in input un aumento dei prezzi di Apple, Microsoft e Starbucks, come suggerito dai dati storici e una significativa diminuzione del prezzo di Tesla. Nonostante ciò abbiamo ottenuto una risposta per il prezzo del Nasdaq crescente, possiamo infatti osservare come la covariata Tesla abbia un coefficiente beta molto basso rispetto agli altri, e quindi influisca poco sulla stima della risposta. E infine osserviamo anche che come da teoria, l’intervallo di previsione venga sempre più ampio di quello di confidenza.