UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M. FANNO"

CORSO DI LAUREA IN ECONOMIA

PROVA FINALE

"ANALISI DI REGRESSIONE PER DATASET AD ALTA DIMENSIONALITÀ"

RELATORE:

CH.MA PROF.SSA ELISA TOSETTI

LAUREANDA: SOFIA CORTESE

MATRICOLA N. 2031870

ANNO ACCADEMICO 2023 – 2024

Dichiaro di aver preso visione del "Regolamento antiplagio" approvato dal Consiglio del Dipartimento di Scienze Economiche e Aziendali e, consapevole delle conseguenze derivanti da dichiarazioni mendaci, dichiaro che il presente lavoro non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Dichiaro inoltre che tutte le fonti utilizzate per la realizzazione del presente lavoro, inclusi i materiali digitali, sono state correttamente citate nel corpo del testo e nella sezione 'Riferimenti bibliografici'.

I hereby declare that I have read and understood the "Anti-plagiarism rules and regulations" approved by the Council of the Department of Economics and Management and I am aware of the consequences of making false statements. I declare that this piece of work has not been previously submitted – either fully or partially – for fulfilling the requirements of an academic degree, whether in Italy or abroad. Furthermore, I declare that the references used for this work – including the digital materials – have been appropriately cited and acknowledged in the text and in the section 'References'.

Firma (signature) ……

Indice

Sommario

Grazie agli sviluppi della tecnologia degli ultimi decenni, i Big Data hanno rivoluzionato il modo in cui le analisi economiche vengono condotte. Questa tesi esplora le caratteristiche principali dei Big Data e la loro importanza in ambito economico. Successivamente, vengono evidenziate le problematiche legate all'utilizzo della regressione multipla, proponendo diverse soluzioni come la Subset Selection, i metodi di Shrinkage e di machine learning. Quest'ultimo prevede diverse tecniche di regressione come la regressione ridge, la regressione lasso ed elastic net.

Utilizzando R, un ambiente di sviluppo open-source per l'analisi statistica, viene analizzata la capacità dei modelli descritti precedemente di gestire grandi volumi di dati e di identificare i fattori statisticamente più rilevanti per la previsione del tasso di crescita del PIL.

Capitolo 1

Big data

1.1 Caratteristiche principali e fonti

Gli avanzamenti nella tecnologia degli ultimi decenni hanno permesso lo sviluppo, l'immagazzinamento e l'utilizzo di una mole molto grande e complessa di dati che provengono da diverse fonti e che riguardano diversi ambiti come la finanza, il marketing, la climatologia, la biologia e la medicina. È possibile evidenziare 4 caratteristiche dei big data, le cosiddette 4 Vs:

- Volume: si riferisce alla quantità di dati che, non solo è grande, ma è anche caratterizzata da una crescita esponenziale. Dal 2010 ci troviamo nell'era degli zettabyte passando da circa 2 zettabyte a più di 100 ZB nel 2023. Il report di IDC prevede che il volume globale dei dati raggiungerà circa 163 ZB nel 2025 e che supererà i 300 ZB nel 2030, Reinsel (2017).

- Velocità: indica la rapidità con cui i dati vengono prodotti ed elaborati. Per esempio, i dati generati in tempo reale, come i messaggi di testo pubblicati su Twitter, vengono registrati nell'architettura dei dati e alcuni microsecondi dopo vengono pubblicati nella cronologia degli utenti. In questo caso, gli utenti generano migliaia di tweet al secondo e i database e i server dell'azienda li gestiscono contemporanemante. In altri casi, i dati sono registrati in tempo reale, ma la loro trasmissione ai server centrali e la loro pubblicazione avviene in un secondo momento. Ad esempio, i dati sulla disoccupazione vengono registrati nel momento in cui una persona aggiorna il proprio stato di occupazione, ma il tasso di disoccupazione viene pubblicato successivamente e in forma aggregata, Kitchin (2016).

- Varietà: fa riferimento a quali formati i dati possono assumere e si distinguono i dati strutturati, i dati semi strutturati e i dati non strutturati, EURONA (2017). I dati strutturati sono formattati in uno schema impostato con campi definiti. Un esempio di dati strutturati è costituito dalle informazioni organizzate all'interno di un foglio di calcolo, dove le colonne rappresentano le variabili e ogni riga corrisponde ai dati relativi a un'unità statistica. Quando più tabelle vengono collegate tra loro grazie a una colonna in comune, ovvero una colonna che contiene i dati della stessa variabile (per esempio il codice fiscale di un utente), viene creato un database relazionale. I dati non strutturati non sono

organizzati e vengono archiviati nel loro formato originale come foto, video, email, messaggi, immagini satellitari e dati di sensori. I dati non strutturati possono avere associati dei dati che hanno una struttura e che quindi vengono definiti semi-strutturati. Per esempio, un'immagine può avere associati dei dati come risoluzione, proprietario, data e ora e dimensioni.

- Veridicità: riguarda la qualità e l'affidabilità dei dati che sono un elemento essenziale per condurre analisi rilevanti e prendere decisioni informate.

Il volume dei big data è così elevato grazie alle nuove e molteplici fonti da cui provengono. Molti dei dati che vengono registrati al giorno d'oggi erano inosservabili fino a qualche decennio fa. È sufficiente pensare all'utilizzo di Internet che gioca un ruolo importante sotto diversi punti di vista:

1. I log creati dai server registrano in modo dettagliato tutte le attività e gli eventi che si verificano nel web, come ricerche di parole chiave, accessi a siti e richieste di log-in. Per esempio, Google Trends mostra la frequenza con cui una parola viene cercata rispetto al volume totale di ricerche. I dati sono disponibili a tutti gli utenti con frequenza settimanale a partire dal 2004 e, considerando il comportamento degli utenti di internet, è evidente che i dati delle ricerche che Google possiede sono una forma di big data. Google Trends è stato utilizzato in diverse aree di ricerca, per esempio Choi and Varian (2012) hanno dimostrato la capacità di previsione della disoccupazione di Google Trends attraverso i report che raccolgono le ricerche di annunci di lavoro. In ambito sanitario si sono distinti Ginsberg et al. (2009) che hanno elaborato un metodo per analizzare le query di ricerca per monitorare le malattie influenzali in una popolazione. Infine, Schmidt and Vosen (2011) grazie alle serie temporali delle query di ricerca, hanno elaborato un indicatore per il consumo privato che supera gli indicatori tradizionali, evidenziando che le informazioni provenienti dai motori di ricerca possono essere una variabile esplicativa del consumo.

2. L'Internet of Things (IoT) consiste in una rete di dispositivi "intelligenti" che sono in grado di raccogliere dati, grazie alla presenza di sensori, e successivamente di scambiarli tramite la loro interconnessione. Per esempio, il contributo di Suryadevara et al. (2013) ha permesso la previsione del comportamento umano utilizzando i dati dei sensori wireless in una smart home. Ulteriori ambiti in cui l'IoT trova applicazione sono l'Industria 4.0, la Smart City, lo Smart Retail, la sanità e l'agricoltura.

3. Le transazioni finanziarie, come i pagamenti elettronici, sono considerati big data grazie all'alta frequenza con cui queste avvengono. Secondo il report World Payments Report di Capgemini (2023), il volume globale delle transazioni elettroniche toccherà 2,3 trilioni di dollari entro il 2027, con un tasso di crescita annuo del 15%. Su scala regionale, i pagamenti digitali aumenteranno del 19,8% nell'area Asia-Pacifico, del 10,7% in Europa e del 6,5% in Nord America.

4. Le piattaforme di e-commerce monitorano gli acquisti e i comportamenti di navigazione degli utenti, immagazzinando informazioni per personalizzare le offerte.

La raccolta dei Big data avviene anche grazie alla rapida evoluzione degli smartphone, permettendo il tracciamento di dati ancora più specifici come l'attività sui social media, il mobile banking, il tracciamento GPS e i dati di altri sensori (microfoni e video). Secondo Deville et al. (2014) i dati degli smartphone possono essere usati per la mappatura della popolazione, producendo stime sulla densità della popolazione a livello spaziale e temporale su scala nazionale.

Tipologie di dataset

In base alla dimensione e alla struttura di un dataset è possibile distinguere tre tipi di dataset. Un "tall" dataset è composto da molte righe (osservazioni) e poche colonne (variabili). Si parla di "Tall" dataset quando i dati variano nel tempo, ad esempio, i "tick-by-tick data": per ogni strumento finanziario vengono registrate tutte le variazioni di prezzo e ogni transazione effettuata in tempo reale. La seconda tipologia è il "fat" dataset che contiene molte colonne (variabili) rispetto al numero di righe (osservazioni). Dentro questa categoria ricade il cross-sectional database che organizza i dati in modo tale da catturare il valore di molte variabili in uno specifico momento e fornire una rappresentazione attuale dell'oggetto di studio. Un esempio può essere un dataset che contiene le voci di bilancio e i principali valori di mercato delle prime dieci aziende che compongono il Dow Jones. Gestire un fat dataset presenta sfide computazionali significative, richiedendo anche di affrontare problemi come la multicollinearità e l'overfitting. I modelli di regressione regolarizzata, come la regressione ridge e la regressione lasso, o i metodi di machine learning permettono di analizzare questi dataset. Infine, un "huge" dataset si distingue per la grande quantità sia di righe (osservazioni) sia di colonne (variabili). Per l'elaborazione e l'archiviazione di questi dataset, sono spesso necessarie infrastrutture di calcolo distribuite che possono essere facilmente scalate aggiungendo nuovi nodi alla rete, permettendo di gestire un numero crescente di dati o di utenti.

1.2 Importanza dei Big Data per l'economia e la finanza

I Big data costituiscono una risorsa fondamentale in tutti gli ambiti, sia per il settore privato sia per il settore pubblico. Le aziende possono sfruttare questo grande volume di dati per ottimizzare i processi decisionali, analizzando le tendenze di mercato e identificando nuove opportunità per sviluppare nuovi prodotti. Inoltre, è possibile migliorare l'efficienza operativa attraverso una riduzione dei costi e lo snellimento dei processi aziendali. Un esempio emblematico dell'ottimizzazione della catena di approviggionamento è Amazon. L'azienda raccoglie enormi quantità di dati su ogni aspetto delle operazioni di acquisto come le abitudini degli utenti, le tendenze di acquisto, i dati demografici, le ricerche sul sito e anche le condizioni meteorologiche. Utilizzando algoritmi di machine learning e analisi predittiva, Amazon è in grado di prevedere la domanda futura di prodotti e mantenere livelli di magazzino ottimali.

Gli istituti bancari utilizzano l'analisi dei big data per identificare e gestire il rischio. Secondo il report di AIFIRM, il rischio di credito[1] risulta essere l'ambito di applicazione con le maggiori opportunità di miglioramento di previsione. Nell'esperimento condotto da uno dei principali Gruppi Bancari italiani nel 2019, sono stati utilizzati tutti i dati bancari disponibili riguardanti i conti correnti e le carte per sviluppare dei modelli di machine learning in grado di identificare i segnali di rischio. Il modello realizzato ha fatto uso di un algoritmo chiamato Extreme Gradient Boosting (XGBoost) che si caratterizza per le sue prestazioni elevate in termini di accuratezza della predizione e velocità di addestramento. I risultati dell'esperimento hanno dimostrato che il modello di machine learning, integrato alle tecniche tradizionali, riesce a cogliere in modo più reattivo i segnali di rischio e a prevedere con maggiore precisione la probabilità di default, AIFIRM (2022).
Infine, per quanto riguarda l'ambito macroeconomico, Aprigliano et al. (2016) hanno utilizzato un modello a fattori dinamici per prevedere la crescita del PIL italiano utilizzando indicatori standard come il consumo, la produzione industriale, l'inflazione, gli indici del mercato azionario e gli indici manifatturieri, e i dati dei sistemi di pagamento come bonifici, assegni e carte. La ricerca ha evidenziato che i dati mensili sui pagamenti aiutano a tracciare il ciclo economico e migliorano il nowcasting. Infine, la regressione lasso ha indicato le variabili del sistema di pagamento come potenziali predittori della crescita del PIL.

1.3 Sfide

L'adozione dei big data comporta numerose sfide che le aziende e le organizzazioni devono affrontare per sfruttare appieno il potenziale di queste tecnologie. Le problematiche dei big data sono legate alle loro caratteristiche distintive (volume, velocità, varietà e veridicità) e riguardano principalmente l'aspetto statistico, organizzativo-informatico e legale, L'Heureux et al. (2017).
A causa del loro volume i Big Data necessitano di infrastutture potenti e scalabili, ossia capaci di gestire un aumento del carico di lavoro o delle dimensioni dei dati mantenendo un buon livello di prestazioni. Un esempio sono i data warehouse distribuiti e il cloud computing. Un ulteriore problema legato alla quantità dei dati è che l'assunzione che i dati siano distribuiti uniformemente tra tutte le classi può venire meno. Questo porta a una sfida nota come *Class Imbalance* la quale può influenzare negativamente le prestazioni di un modello. Infine, un altro fenomeno che si afferma man mano che aumenta il volume dei dati è quello dell'overfitting: il modello performa bene sul dataset di addestramento ma non riesce a generalizzare adeguatamente per nuovi dati.
Per quanto riguarda la varietà, una delle principali sfide è l'analisi di dati eterogenei. Nei dati si possono riconoscere due categorie di eterogeneità: l'eterogeneità sintattica e semantica. L'eterogeneità sintattica si riferisce alla diversità nei tipi di dati, nei formati dei file, nella codifica dei

[1] Def: Il rischio di credito rappresenta la possibilità che il debitore non adempia ai propri obblighi contrattuali di rimborso del capitale e/o degli interessi

dati. L'eterogeneità semantica, invece, si riferisce alle differenze nei significati e nelle interpretazioni. Per eseguire delle analisi con dataset che integrano diversi tipi di dati, queste variazioni sintattiche e semantiche devono essere gestite adeguatamente.
Rispetto alla veridicità dei dati, è necessario sottolineare l'importanza di tracciare la provenienza dei dati e i loro spostamenti e di assicurare la loro integrità e qualità per evitare di incorrere in dati incompleti o inaffidabili. Inoltre, i big data spesso contengono informazioni personali e il loro utilizzo solleva questioni etiche significative. Di conseguenza, gli utenti devono essere informati in merito a come questi dati vengono raccolti ed utilizzati e devono poter contare sulla loro protezione. In questo contesto, la Commissione Europea ha approvato il GDPR (Regolamento Generale sulla Protezione dei Dati) che stabilisce requisiti rigorosi per la protezione dei dati personali.
E' possibile concludere che per affrontare queste sfide è necessario un approccio globale che combini innovazione tecnologica e competenze umane in grado interpretare correttamente i dati e di tradurli in informazioni.

Capitolo 2

Modelli di regressione per i dataset di grandi dimensioni

2.1 Il modello OLS e le sue problematiche

Le tecniche tradizionali di regressione si prestano perfettamente per dataset di piccole dimensioni nei quali il numero delle osservazioni è maggiore del numero delle variabili. Il metodo Ordinary Least Square, minimizzando la somma dei quadrati dei residui, produce le stime dei coefficienti del seguente modello di regressione, che esprime la relazione tra la variabile dipendente y e un set di variabili $x_1, x_2, \ldots, x_p$. Attraverso:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \tag{2.1}$$

Vengono prodotte le stime dei coefficienti della regressione:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \tag{2.2}$$

Secondo il teorema di Gauss-Markov, le assunzioni che devono essere rispettate affinché gli stimatori dei coefficienti della regressione OLS siano i migliori stimatori lineari non distorti a varianza minima ("The Best Linear Unbiased Estimators") sono:

- Il modello è lineare nei parametri, cioè la relazione tra la variabile dipendente e le variabili esplicative è espressa come una combinazione lineare dei parametri.
- Assenza di correlazione tra le variabili esplicative e il termine d'errore, quindi $\mathbb{E} = [\varepsilon_i | X_i] = 0$. Se all'interno del termine d'errore ricade una variabile correlata con le variabili già inserite nel modello e determinante per la variabile indipendente, allora si verifica distorsione da variabile omessa.
- La presenza di omoschedasticità garantisce che la varianza dell'errore sia costante al variare delle variabili esplicative. Se gli errori sono eteroschedastici le stime dei coefficienti rimangono corrette e consistenti ma la loro varianza aumenta.

- L'assenza di autocorrelazione degli errori prevede che questi non siano correlati tra loro, quindi il valore dell'errore per una osservazione non deve fornire alcuna informazione sul valore dell'errore per un'altra osservazione.

Con i dataset di grandi dimensioni che si sono sviluppati negli ultimi decenni, in particolare i "fat" dataset che hanno un numero di variabili maggiore rispetto al numero di unità statistiche, costruire un modello econometrico che include tutti i regressori può portare all'overfitting o sovradattamento: il modello si adatta perfettamente ai dati di addestramento (training data) ma non è in grado di effettuare previsioni altrettanto accurate con nuovi dati (testing data). Se aumenta la sommatoria dei residui al quadrato perché i dati non sono ben approssimati dal modello di regressione, allora aumenta la varianza delle stime dei coefficienti in base alla seguente formula:

$$Var(\hat{\beta}_j) = \frac{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.3}$$

Inoltre, l'inclusione di variabili non rilevanti all'interno di una regressione contraddice il principio di parsimonia secondo il quale, in presenza di più modelli che descrivono efficacemente il dataset considerato, è bene preferire quello più semplice, ossia quello con un numero minore di variabili. È stato quindi indispensabile creare modelli econometrici capaci di adattarsi ai nuovi tipi di dataset, al fine di soddisfare due requisiti fondamentali: l'accuratezza della predizione e l'interpretazione del modello.

1. L'accuratezza della predizione è la capacità del modello di fare previsioni sui dati non utilizzati per la stima del modello, i testing data. Con i dataset di piccole dimensioni, la regressione OLS produce stime che avranno una bassa varianza e quindi l'accuratezza della predizione sarà elevata. Come già illustrato, l'utilizzo di dataset di grandi dimensioni comporta una varianza significativamente maggiore nelle stime dei coefficienti OLS, causando problemi sia nella capacità predittiva del modello che nell'inferenza. Attraverso la selezione di alcune variabili, è possibile ridurre sensibilmente la varianza a discapito di un piccolo aumento del bias dei coefficienti. Questa caratteristica viene chiamata bias-variance trade off. Una delle statistiche che verranno analizzate per determinare l'accuratezza predittiva è il Root Mean Squared Error dato dalla radice della media degli errori al quadrato.

2. L'interpretazione del modello diventa difficile quando molte variabili vengono incluse nel modello perché alcune potrebbero non essere correlate con la variabile dipendente. Inoltre, alcune variabili esplicative potrebbero presentare un'elevata correlazione tra loro, causando multicollinearità. Questo fenomeno non consente di isolare l'effetto che ciascuna variabile esplicativa ha sulla variabile dipendente. Si può concludere che rimuovendo alcuni regressori si può ottenere un modello di più facile interpretazione. Tuttavia, il metodo OLS non tende a ridurre le stime dei coefficienti esattamente a zero, rendendo necessarie altre procedure di selezione automatica.

2.2 I metodi di selezione delle variabili

Esistono diversi metodi per la selezione delle variabili più importanti all'interno di un dataset come Subset Selection e Shrinkage, che verranno descritti di seguito.

2.2.1 Subset Selection e criteri di selezione

Subset Selection è una tecnica che permette di identificare un sottoinsieme di variabili che tra tutte risultano essere le più rilevanti. È possibile distinguere diversi procedimenti:

1. Forward Selection
2. Backward Selection
3. Stepwise Selection

Forward selection considera inizialmente un modello M_0 che contiene solo l'intercetta e ad ogni step aggiunge una variabile. Quindi, in presenza di k variabili, vengono stimate k regressioni. Per lo step successivo si sceglie di proseguire con il modello che contiene la variabile che restituisce il miglior modello (M_1). Allo step successivo, verranno stimate k – 1 regressioni nelle quali verrà inclusa un'altra variabile al modello M1 e si sceglie di proseguire con il modello migliore tra quelli presenti (M_2). Si continua con questo procedimento fino all'ultimo step quando il modello Mk contiene tutti i regressori.
Backward Selection procede a ritroso, considerando inizialmente il modello che contiene tutte le variabili. Allo step successivo vengono stimate k regressioni, ciascuna con una variabile in meno, e si identifica la variabile meno significativa, che verrà rimossa dal modello. Si continua fino a quando non si ottiene modello migliore nel quale la rimozione di una delle variabili comporta un peggioramento significativo della bontà di adattamento.
Infine, Stepwise Selection, è una procedura ibrida che combina forward selection e backward selection, dove le variabili possono essere aggiunte o rimosse ad ogni step.
Per tutti i procedimenti descritti, il criterio da considerare per selezionare ad ogni step il miglior modello può essere l'R^2 corretto, l'AIC o il BIC. L'R^2 misura la proporzione della varianza della variabile dipendente che è spiegata dal modello. Tuttavia, R^2 tende ad aumentare man mano che si aggiungono variabili al modello anche se queste non sono significative. Di conseguenza, l'R^2 non può essere utilizzato come criterio di selezione ed è necessario effettuare un aggiustamento penalizzando l'inclusione di più variabili come segue:

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{RSS}{TSS} \tag{2.4}$$

L'aggiunta di un regressore ha due effetti sull'R^2corretto: L'RSS diminuisce ma il fattore $\left(\frac{n-1}{n-k-1}\right)$ aumenta, quindi il risultato finale dipende da quale dei due effetti è maggiore. L'R^2 corretto risulta essere minore rispetto all'R^2 anche se per dimensionalità elevate le due statistiche tendono ad eguagliarsi.

L'AIC e il BIC sono due criteri di informazione composti dalla seguente struttura:

$$IC = f(RSS) + g(k) \tag{2.5}$$

Una funzione f che considera l'RSS e una funzione g che considera il numero di variabili esplicative k. L'obiettivo è quello di ottenere l'RSS più basso possibile, ma per fare ciò è necessario aumentare il numero di variabili. I seguenti criteri cercano di trovare un compromesso tra RSS e numero di parametri per ottenere un modello che raggiunga un buon adattamento ai dati senza essere troppo complesso.
L'AIC (Akaike information criterion) è dato da:

$$AIC = \log\left(\frac{RSS}{n}\right) + \frac{2k}{n} \tag{2.6}$$

Greene (2002). La statistica AIC comprende un termine di penalità $\frac{2k}{n}$ che aumenta man mano che si inseriscono variabili esplicative in modo da penalizzare i modelli poco parsimoniosi. Durante le fasi di subset selection la scelta ricade sul modello con AIC più basso.
Il BIC (Bayesian information criterion) è dato da:

$$BIC = \log\left(\frac{RSS}{n}\right) + \frac{k\log(n)}{n} \tag{2.7}$$

Greene (2002). La formulazione è simile a quella dell'AIC, ma il termine di penalità è più severo perchè il 2 è sostituito da $log(n)$ che è maggiore di 2 per una numerosità campionaria sufficientemente elevata. È necessario ricordare che l'obiettivo non è tanto raggiungere un determinato livello di AIC o BIC, piuttosto riuscire a confrontare la stessa statistica per modelli diversi in modo da determinare il migliore.

2.2.2 Shrinkage

Shrinkage, o regolarizzazione, è un metodo che consente di stimare un modello che contiene tutte le variabili presenti nel dataset, ma che riduce le stime di alcuni coefficienti quasi o esattamente a zero. Esistono diversi tipi di regolarizzazione: Regressione Ridge, Regressione Lasso ed Elastic Net.

Regressione Ridge

La regressione Ridge prevede la minimizzazione della seguente formula:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.8}$$

T. Hastie and Friedman (2001). Il punto di partenza per costruire la regressione Ridge è sempre lo stesso: trovare le stime dei coefficienti che minimizzano della sommatoria dei residui al quadrato alla quale viene aggiunto un ulteriore termine: $\lambda \sum_{j=1}^{p} \beta_j^2$, ossia la sommatoria dei

coefficienti al quadrato moltiplicata per un coefficiente di penalizzazione λ detto parametro di tuning. Questo termine $\lambda \sum_{j=1}^{p} \beta_j^2$, chiamato penalità shrinkage, può essere considerato come un vincolo da rispettare nella minimizzazione della sommatoria dei residui al quadrato.
È evidente l'esistenza di un trade off tra la minimizzazione dell'RSS e della penalità shrinkage: l'RSS viene minimizzato quando il modello di regressione contiene tutti i parametri in quanto i dati riescono ad adattarsi perfettamente al modello, mentre la penalità shrinkage viene minimizzata quando la sommatoria dei coefficienti è vicina allo zero. La risoluzione del problema prevede di far tendere a zero un numero sufficiente di coefficienti in modo tale da ridurre la varianza del modello e raggiungere una buona bontà di adattamento.
Il parametro di tuning può assumere solo valori non negativi ($\lambda \geq 0$) e determina l'impatto della penalità di shrinkage sulle stime dei coefficienti. Man mano che il valore di lambda aumenta, l'impatto che la minimizzazione della sommatoria dei coefficienti al quadrato ha sulle stime dei coefficienti è sempre maggiore e questi tenderanno a zero. Quando $\lambda = 0$, la penalità shrinkage perde completamente il suo effetto, quindi la regressione Ridge produrrà le stesse stime della regressione OLS. Per ogni livello di lambda ci saranno diversi set di stime dei coefficienti; infatti, selezionare il lambda corretto è il punto cruciale per produrre un buon modello.
La penaltà shrinkage è applicata solo a $\beta_1, \ldots, \beta_p$ e non all'intercetta β_0 dato che questa rappresenta solo il valore atteso che la variabile y assume quando tutte le variabili dipendenti sono pari a zero.

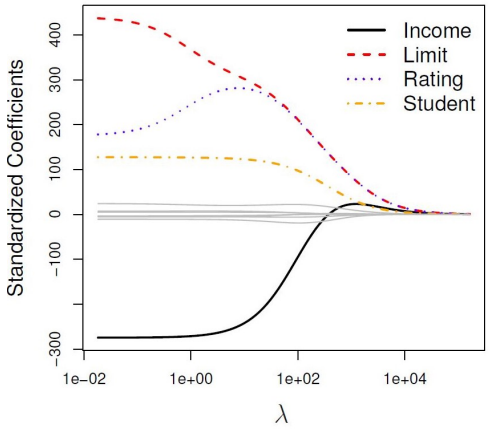


Figura 2.1: In questo grafico è possibile notare l'andamento delle stime dei coefficienti delle variabili elencate in funzione del parametro di tuning, G. James and Tibshirani (2013).

Vantaggi e svantaggi della Regressione Ridge

Il principale vantaggio della regressione Ridge rispetto alla regressione OLS risiede nella riduzione dell'overfitting grazie all'introduzione del termine di penalizzazione. Man mano che il parametro di tuning aumenta, un numero sempre maggiore di coefficienti convergerà allo zero, portando a una diminuzione sostanziale della varianza a costo di un leggero aumento del bias.

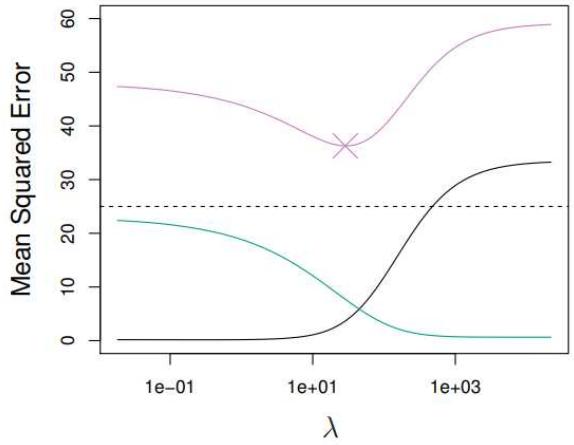


Figura 2.2: Nel grafico vengono rappresentati l'andamento della varianza (curva verde), l'andamento del bias (curva nera) e l'andamento del test mean squared error (rosa) della regressione ridge al variare del parametro di tuning. È possibile notare che la varianza diminuisce considerevolmente a costo di un leggero incremento nel bias dei coefficienti, per cui è necessario trovare il valore di lambda tale per cui sia il bias che la varianza siano minori possibili, considerando che hanno tendenza opposta. Infine, in corrispondenza del punto di incontro tra varianza e bias, il mean squared error raggiunge il suo punto di minimo, G. James and Tibshirani (2013).

Lo svantaggio principale della Regressione Ridge è l'inclusione di tutte le variabili all'interno del modello. Il termine di penalità fa tendere alcuni coefficienti a zero man mano che aumenta lambda, ma nessuno di questi raggiunge esattamente lo zero. Di conseguenza, anche se l'impatto di una o più variabili può essere ridotto, nessuna verrà mai esclusa dal modello. Questo non è un problema per l'accuratezza del modello ma potrebbe creare problemi nella sua interpretazione nel caso in cui il numero delle variabili sia particolarmente alto.

Regressione Lasso

La regressione Lasso (Least Absolute Shrinkage and Selection Operator) prevede la minimizzazione della seguente formula:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.9}$$

T. Hastie and Friedman (2001). La regressione Lasso supera lo svantaggio della regressione Ridge sostituendo al termine $\sum_{j=1}^{p} \beta_j^2$ il termine $\sum_{j=1}^{p} |\beta_j|$.
In questo caso il termine di penalità ha l'effetto di stimare alcuni coefficienti esattamente uguali a zero, eliminando delle variabili dal modello e svolgendo quindi una vera e propria selezione quando il parametro di tuning raggiunge un valore sufficiente.

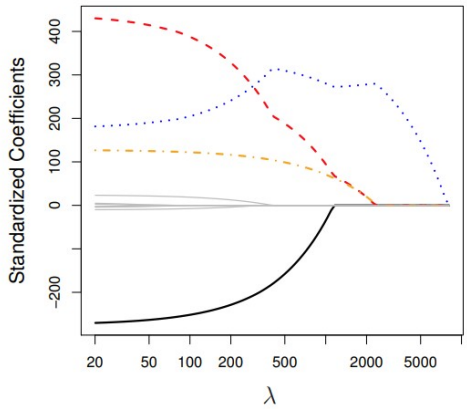


Figura 2.3: In questo grafico è possibile notare come, all'aumentare del valore di lambda, cambino le stime dei coefficienti fino ad arrivare, una alla volta, esattamente pari a zero, G. James and Tibshirani (2013).

La regressione Lasso, vincolando alcuni coefficienti allo zero, può causare un aumento significativo nel bias delle stime dei coefficienti. Utilizzare queste stime per condurre inferenza sui parametri può risultare problematico. Per ovviare a questo svantaggio, è possibile utilizzare la regressione Post – Lasso che combina i punti di forza della Regressione Lasso e della Regressione OLS per eliminare il bias. Per ottenere le stime dei coefficienti della regressione Post-Lasso è necessario minimizzare il seguente problema:

$$min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \text{ sotto il vincolo } \beta_j = 0 \text{ se } \widetilde{\beta}_j = 0 \tag{2.10}$$

Dove $\widetilde{\beta}_j$ è la stima del coefficiente effettuata tramite la regressione Lasso, Belloni and Chernozhukov (2013).
Quindi, la regressione Post-Lasso si svolge in due fasi: Alla prima fase si utilizza la regressione Lasso per selezionare il sottoinsieme di variabili rilevanti ai fini della predizione. Alla seconda fase si esegue una regressione OLS sulle variabili selezionate precedentemente.

Confronto tra Regressione Ridge e Regressione Lasso

La differenza sostanziale risiede nell'assunzione iniziale con la quale si costruisce il modello. La regressione Ridge viene stimata ipotizzando *approximate sparsity* secondo la quale alcuni dei coefficienti sono approssimativamente zero e l'errore di approssimazione è minimo. L'assunzione per costruire dei modelli che non includano tutti i regressori, come la regressione Lasso, è detta *sparsity*. Secondo questa assunzione, solo un gruppo di variabili è importante per la predizione accurata della variabile dipendente. Le stime degli altri coefficienti vengono stimate esattamente pari a zero, anche se è possibile sapere quali. La regressione Lasso svolge una sorta di variable selection dato che seleziona solo un set di variabili. Inoltre, è possibile evidenziare la differenza tra Regressione Ridge e Regressione Lasso formulando dei problemi di ottimizzazione vincolata:

Per la regressione Ridge:

$$\text{minimizzare} \quad \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \quad \text{sotto il vincolo} \quad \sum_{j=1}^{p}\beta_j^2 \leq s \qquad (2.11)$$

Per la regressione Lasso:

$$\text{minimizzare} \quad \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \quad \text{sotto il vincolo} \quad \sum_{j=1}^{p}|\beta_j| \leq s \qquad (2.12)$$

In questi problemi di ottimizzazione è possibile paragonare la minimizzazione dell'RSS alla minimizzazione di una funzione di costo, la quale deve rispettare un "budget". Il budget, in questo caso, è dato dalla sommatoria dei coefficienti che non deve superare un certo valore s per il quale esiste una corrispondenza univoca con il valore del parametro lambda. Tramite la formulazione di questo problema e la rappresentazione grafica è possibile comprendere perché la regressione Lasso è in grado di stimare alcuni coefficienti esattamente uguali a zero, al contrario della regressione ridge.

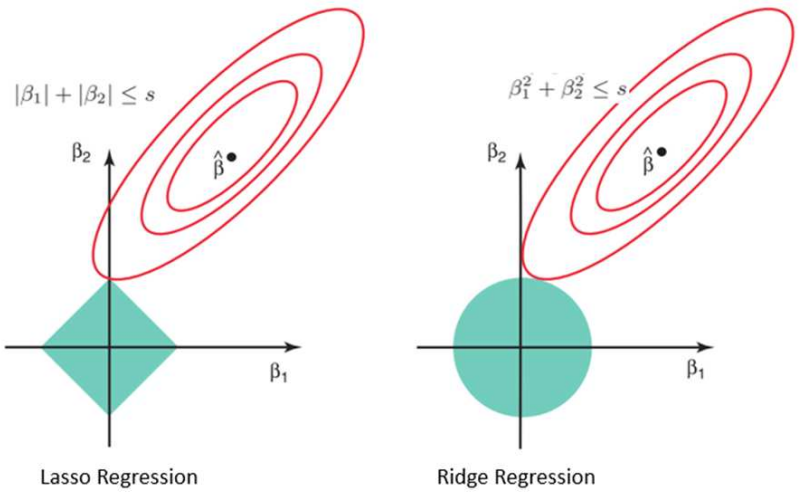


Figura 2.4: Per semplicità viene analizzato il caso bidimensionale, tuttavia le stesse conclusioni possono essere tratte anche quando il numero di parametri aumenta. In entrambi i grafici, gli assi cartesiani rappresentano i coefficienti della regressione e ogni ellisse rappresenta il valore dell'RSS che si ottiene in corrispondenza delle stime in quel punto. Man mano che si allontana dalle stime OLS $\hat{\beta}$, l'RSS aumenta.
Le stime dei coefficienti della regressione Ridge si ottengono nel punto in cui l'ellisse è tangente al vincolo rappresentato dal cerchio. Dato questo vincolo circolare, il punto di tangenza non sarà in corrispondenza di un'asse cartesiano, quindi le stime dei coefficienti della regressione Ridge non potranno essere esattamente zero. Per quanto riguarda la regressione Lasso, invece, le stime dei coefficienti si ottengono nel punto in cui l'ellisse è tangente al vincolo rappresentato dal quadrato, dunque il punto di tangenza può essere in corrispondenza di un asse cartesiano dove un coefficiente viene stimato esattamente uguale a zero, T. Hastie and Friedman (2001).

Si può generalizzare la regressione Ridge e Lasso tramite la seguente formula:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \tag{2.13}$$

Dove il variare di q determina la forma del vincolo e per valori di q compresi tra 1 e 2 è possibile giungere ad un compromesso tra la regressione Ridge e la regressione Lasso.

Elastic Net

Elastic Net è una tecnica di regressione che combina le proprietà della regressione Ridge e della regressione Lasso per migliorare le prestazioni predittive e la selezione delle variabili. È utile quando ci sono molte variabili correlate perché in questo caso la regressione Lasso risulta essere instabile. In particolare, se un gruppo di variabili sono altamente correlate tra loro, la regressione Lasso tende a selezionarne solo una senza dare importanza a quale specifica variabile viene scelta. La regressione Ridge, invece, in presenza di variabili correlate tra loro, tende a portare i coefficienti di queste variabili uno verso l'altro. Il termine di penalità dell'Elastic net diventa:

$$\lambda \sum_{j=1}^{p} (\alpha|\beta_j| + (1-\alpha)\beta_j^2) \tag{2.14}$$

T. Hastie and Friedman (2001), dove il parametro α bilancia l'importanza della penalizzazione Lasso e della penalizzazione Ridge:

- Se $\alpha = 0$, l'Elastic Net diventa una regressione Ridge
- Se $\alpha = 1$, l'Elastic Net diventa una regressione Lasso
- Se $\alpha \in (0,1)$: l'Elastic Net combina entrambe le penalizzazioni.

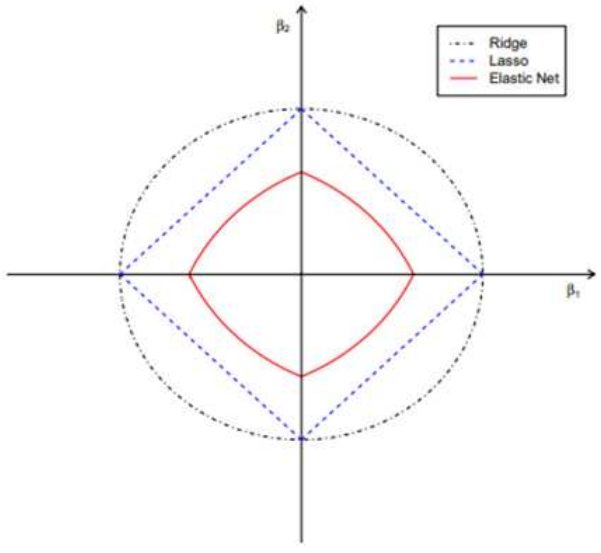


Figura 2.5: Confronto tra la penalità della regressione Ridge (linea nera), regressione Lasso (linea azzurra) ed Elastic Net per un valore di alpha pari a 0.5, T. Hastie and Friedman (2001).

Fused Lasso

In molte applicazioni reali può accadere che in un dataset l'insieme delle variabili abbiano una struttura di sequenzialità. Per esempio, i dati finanziari ed economici come i prezzi dei titoli azionari, la variazione dei tassi di interesse o il valore del PIL sono serie storiche nelle quali i dati sono ordinati nel tempo (giorni, trimestri o anni). Per questa tipologia di dataset, oltre a selezionare le variabili statisticamente rilevanti, è necessario imporre un ordinamento preciso in modo da mantenere la continuità temporale tra i coefficienti di variabili vicine nel tempo.
Per ottenere questo risultato, viene utilizzata la regressione Fused Lasso:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j| \tag{2.15}$$

T. Hastie and Friedman (2001). La funzione da minimizzare è composta dall'RSS, dalla sommatoria dei coefficienti in valore assoluto con il relativo parametro di penalizzazione λ_1 e dalla sommatoria delle differenze tra coefficienti consecutivi con il termine di penalizzazione λ_2.
Penalizzare le differenze tra coefficienti consecutivi favorisce soluzioni in cui i coefficienti adiacenti sono simili in modo da migliorare l'interpretabilità del modello. Per esempio, ci si aspetta che variabili vicine nel tempo abbiano un impatto simile sulla variabile dipendente e, inoltre, i coefficienti dovrebbero cambiare gradualmente nel tempo anziché drasticamente.

2.2.3 La scelta del parametro di tuning

Per creare un modello di regressione come i precedenti è necessario selezionare un valore adatto per il parametro di tuning λ, definito anche come iperparametro di regolarizzazione. Un iperparametro è un parametro esterno al modello che controlla la sua struttura e che non può essere stimato a partire dai dati. La tecnica più utilizzata per determinare il valore di λ è la cross validation. Esistono diversi metodi di cross validation, uno di questi è il K-Fold CV. Il K-Fold CV consiste nel dividere il dataset in k sottogruppi della stessa dimensione, chiamati fold, che vengono utilizzati per testare il modello di regressione con lo scopo di migliorare l'accuratezza e, allo stesso tempo, evitare l'overfitting.
Il primo passo per determinare il parametro λ ottimo è definire una griglia di valori per lambda. Successivamente, per ognuno di questi valori, il modello di regressione viene allenato k volte, utilizzando k-1 fold per l'allenamento e il rimanente per la validazione. Per ogni fold di validazione viene calcolato l'MSE e alla fine, quando il processo è stato ripetuto k volte, si calcola una media per ottenere il cross-validation error.
La configurazione ottimale del modello consiste nell'identificare l'iperparametro λ che ha prodotto il minor cross-validation error.
Infine, il modello è stimato nuovamente utilizzando il valore di lambda ottimo e l'intero dataset di addestramento, per poi valutare le prestazioni con un test dataset separato.
Si può concludere affermando che la cross validation è un metodo di valutazione delle prestazioni di un modello affidabile e accurato perché prevede la suddivisione in più sottogruppi

con la successiva aggregazione dei risultati. Tale approccio contribuisce a ridurre la varianza nelle stime delle prestazioni. Inoltre, la cross validation permette il massimo utilizzo dei dati a disposizione perchè in ogni step una parte diversa del dataset viene utilizzata come dataset di validazione, mentre i rimanenti fold vengono utilizzati per l'addestramento. Questo procedimento è utile soprattutto quando il numero di unità statistiche a disposizione è ridotto.

2.2.4 Confronto dei modelli: RMSE

L'RMSE (Root Mean Square Error) o errore quadratico medio è una delle metriche più utilizzate per confrontare modelli di regressione ed è dato da:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \tag{2.16}$$

Dove $(y_i - \hat{y}_i)$ rappresenta la differenza tra valori osservati e valori predetti, dunque i residui. Grazie alla sua formula, l'RMSE considera:

- La distorsione: la differenza tra il valore osservato e i valori predetti.
- La varianza: la dispersione delle previsioni attorno alla media.

Questo bilanciamento è importante per valutare quanto un modello sia ben adattato ai dati, evitando l'overfitting.
Prima di calcolare la media e la radice quadrata, si elevano al quadrato gli errori e ciò consente di penalizzare di più gli errori elevati rispetto a quelli più piccoli, a differenza del MAE (Mean Absolute Error)[1]. Un'altra caratteristica dell'RMSE è che, grazie alla radice quadrata, mantiene la stessa unità di misura della variabile dipendente, quindi è più facilmente confrontabile, a differenza dell'MSE (Mean Square Error) [2].
Inoltre, l'RMSE viene utilizzato nella Regressione Ridge e nella Regressione Lasso come misura da minimizzare grazie alle sue proprietà di continuità e differenziabilità. La funzione del MAE, a causa della presenza del valore assoluto, non è differenziabile nel punto in cui la differenza tra valori osservati e predetti è zero.

2.3 Machine Learning

Il Machine Learning è un'area dell'Intelligenza Artificiale che permette di creare modelli che apprendono dai dati presentati e successivamente effettuano previsioni sui nuovi dati. Esistono diverse tecniche di Machine Learning come:

- Apprendimento Supervisionato: Il modello viene addestrato su un dataset etichettato con l'obiettivo di trovare una funzione che spieghi la variabile dipendente a partire dalle varia-

[1] Dato dalla media del valore assoluto dei residui
[2] Dato dalla media dei residui al quadrato

bili esplicative. L'apprendimento automatico include, per esempio, la regressione lineare, la regressione logistica, gli alberi decisionali e gli ensemble methods.

- Apprendimento Non Supervisionato: I dati presenti non sono etichettati e l'obiettivo è scoprire strutture nascoste o pattern nei dati. Le tecniche più comuni sono il clustering e la riduzione della dimensionalità.
- Apprendimento Rinforzato: Un agente esplora l'ambiente e riceve ricompense o penalità in base alle azioni intraprese, con l'obiettivo di massimizzare la ricompensa cumulativa nel lungo termine.

2.3.1 Gli alberi decisionali

Gli alberi decisionali costituiscono una tecnica di apprendimento supervisionato utilizzata per problemi di regressione e classificazione. Questa tecnica prevede la suddivisione dello spazio dei predittori in un numero di regioni semplici, ossia in sottogruppi omogenei caratterizzati da variabili di risposta simili. I sottogruppi vengono formati in modo ricorsivo utilizzando suddivisioni binarie. Questo processo viene ripetuto più volte fino a quando non si soddisfa una regola di arresto adeguata, per esempio un numero massimo di nodi o una specifica soglia di RSS.

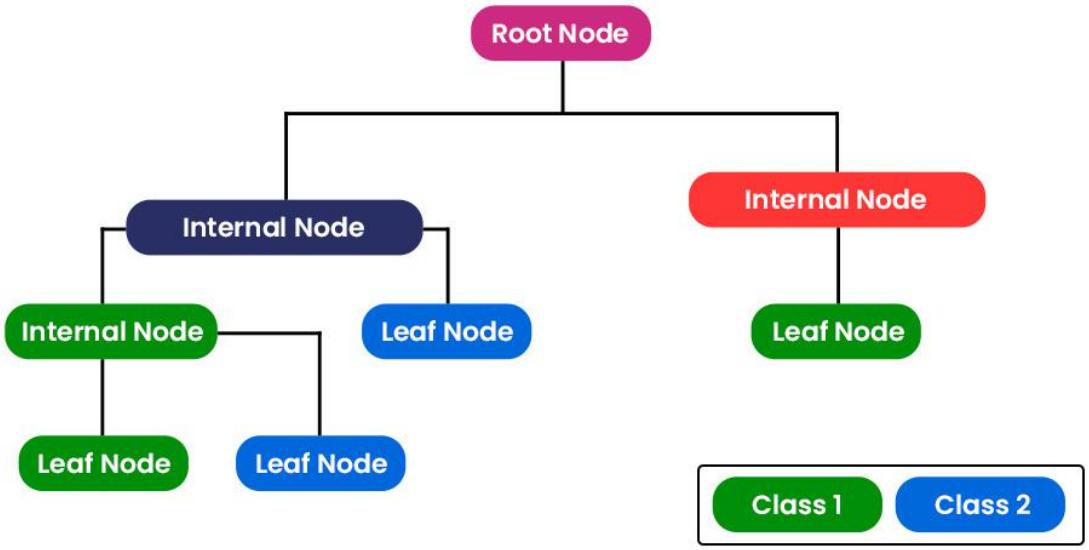


Figura 2.6: In questa immagine viene mostrata la struttura di un albero decisionale, dove ogni leaf node corrisponde a una regione.

Come vengono costruite le regioni: Ipotizziamo di avere k regressori. Lo spazio dei predittori, cioè l'insieme dei valori possibili per $X_1, X_2, X_3, ..., X_k$, viene diviso in J regioni distinte e non sovrapposte $(R_1, R_2, ..., R_J)$ minimizzando la sommatoria dei residui al quadrato data da:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j}) \tag{2.17}$$

dove $\hat{y}_{R_j}$ è la media della variabile dipendente all'interno della j-esima regione, G. James and Tibshirani (2013).

Metodi di ensemble Gli alberi decisionali possono facilmente adattarsi troppo ai dati di addestramento, portando all'overfitting, tuttavia questo problema è facilmente risolvibile combinando molti alberi di regressione insieme in un ensemble complessivo.
Il Bagging è un metodo di ensemble che consiste nell'estrarre diversi campioni casuali dal dataset di addestramento originale, creare un albero decisionale per ogni campione e infine calcolare la media delle previsioni ottenute dai singoli modelli.

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x) \tag{2.18}$$

dove $\hat{f}^b(x)$ è la previsione del modello addestrato sul campione b-esimo ottenuto tramite il metodo bootstrap G. James and Tibshirani (2013). Considerando la media delle previsioni ottenute dai diversi campioni si riduce la varianza poichè la varianza della media è data da $\frac{\sigma^2}{n}$ che è minore della varianza σ^2 di un singolo campione.
Il Gradient Boosting funziona in modo simile, tuttavia gli alberi vengono creati in sequenza, utilizzando le informazioni degli alberi creati precedentemente.

- Si crea un albero decisionale $Y = f_1(X)$ e si calcolano i residui r_1.
- Si crea un albero decisionale utilizzando come variabile dipendente i residui r_1 e si calcolano nuovamente i residui r_2.
- Si crea un albero decisionale per i residui r_2 e il processo continua finchè una regola di arresto non viene soddisfatta.

Ad ogni step i residui vengono aggiornati, permettendo al modello di migliorare lentamente nelle aree in cui non performa bene. Inoltre, ad ogni albero decisionale stimato sui residui viene assegnato un peso chiamato Learning Rate in base al quale il modello impara più o meno velocemente.

Capitolo 3

Applicazione con R al dataset Penn World Table

3.1 Introduzione

In questo capitolo, grazie al database Penn World Table, verrà svolta un'applicazione dei metodi di selezione delle variabili e delle tecniche di regressione ridge e lasso per enfatizzare le differenze tra le due. In primo luogo verrà prodotto un modello di regressione OLS e verranno evidenziate le sue problematiche. Successivamente verrà utilizzata la stepwise selection per la selezione delle variabili con due criteri di selezione, R^2 corretto e BIC. In seguito, verranno generati e paragonati il modello di regressione ridge e il modello di regressione lasso per il quale verranno prodotte stime corrette e consistenti grazie alla regressione post-lasso. Per concludere, le variabili del modello verranno analizzate per spiegare come, secondo la teoria macroeconomica, queste influiscano sul tasso di crescita del Pil.

3.2 Caratteristiche del dataset

La Penn World Table è un database sviluppato da Feenstra (Università della California) e da Inklaar e Timmer (Università di Groningen). Questo dataset contiene una serie di dati economici riguardo a 183 paesi, tra il 1950 e il 2019, che forniscono informazioni dettagliate e comparabili nel tempo, grazie alla normalizzazione ad un anno base (2017), e tra paesi dato che i dati sono espressi in dollari e anche tramite l'indice PPA. Le variabili del Penn World Table sono quarantuno e possono essere raggruppate come segue:

1. Dati riguardanti il PIL (Prodotto Interno Lordo) espresso secondo l'indice PPA e a prezzi correnti e a prezzi costanti per eliminare gli effetti dell'inflazione.

2. Dati sulla popolazione ed occupazione

3. Dati dettagliati sulle componenti del PIL come consumo, investimenti, esportazioni ed importazioni.

4. Dati sulla produttività totale dei fattori (PTF) che misura l'efficienza produttiva.

5. Dati sui livelli dei prezzi rispetto agli Stati Uniti d'America.

Il Penn World Table utilizza l'indice PPA (Parità di Potere d'Acquisto) per confrontare i dati tra paesi, eliminando le distorsioni causate dalle fluttuazioni dei tassi di cambio e dalle differenze di prezzo per i beni di consumo. L'idea alla base della PPA è che un paniere di beni identici dovrebbe avere lo stesso prezzo in due paesi quando espresso in una comune unità monetaria. Se non è così, esiste un'opportunità di arbitraggio che, teoricamente, dovrebbe essere eliminata dalle forze di mercato nel lungo periodo.
Dato il vettore di prezzi relativo ad un paese (p_i) e il vettore di prezzi relativo al paese di riferimento (π_i), in questo caso gli Stati Uniti, il tasso della PPA può essere definito come:

$$PPA_i = \frac{p_i * q_i}{\pi_i * q_i} \tag{3.1}$$

Il tasso di PPA viene impiegato, per esempio, nella seguente formula per calcolare la variabile $CGDP^e$(Converted Gross Domestic Product):

$$CGDPE_i^e = \frac{p_i q_i}{PPP_i} + \frac{(X_i - M_i)}{PPP_i} = \frac{GDP_i}{PPP_i} \tag{3.2}$$

Sostanzialmente, si considera la spesa totale in beni e servizi nazionale "aggiustata" dal tasso di PPA e si aggiunge il saldo commerciale anch'esso "aggiustato" dal tasso di PPA, Robert C. Feenstra and Timmer (2015).

3.3 Codici e risultati

Per questa applicazione in R Studio sono stati scaricati i dati di 57 paesi nell'anno 2019. L'obiettivo è quello di analizzare quali sono le variabili più statisticamente rilavanti per la predizione del tasso di crescita del PIL, quindi è stato calcolato, per ogni paese, il tasso di crescita annuale del PIL tra il 2018 e il 2019. Il tasso di crescita del PIL (gr) viene definito come differenza tra $log(GDP_t)$ e $log(GDP_{t-1})$.

```
PWT <- read_excel("pwtpaesi.xlsx", sheet="mydata")
PWT$rgdpe18_l <- log(PWT$rgdpe18)
PWT$rgdpe19_l <- log(PWT$rgdpe19)
PWT$gr <- PWT$rgdpe19_l - PWT$rgdpe18_l
```

La variabile scelta per il calcolo del tasso di crescita del PIL è rgdpe che, essendo calcolata con l'indice PPA, permette il confronto tra più paesi.
Di seguito vengono prodotte le pricipali statistiche della variabile dipendente gr e gli indici di simmetria e curtosi in modo da definire la sua distribuzione.

```
mean(PWT$gr)
median(PWT$gr)
skewness(PWT$gr)
kurtosis(PWT$gr)
ggplot(PWT) + aes(gr) + geom_density(fill="grey") + theme_bw()
```

```
## [1] 0.02157026
## [1] 0.01782909
## [1] 1.6499
## [1] 9.878769
```

La media è maggiore della mediana, quindi la distribuzione della variabile dipendente presenta valori più estremi nella coda di destra. Il valore dell'indice di skewness lo conferma, dato che è maggiore di zero. Infine, l'indice di Curtosi è ampiamente maggiore di zero, ciò significa che la distribuzione è leptocurtica, con code più pesanti e picco più alto rispetto alla distribuzione Normale.

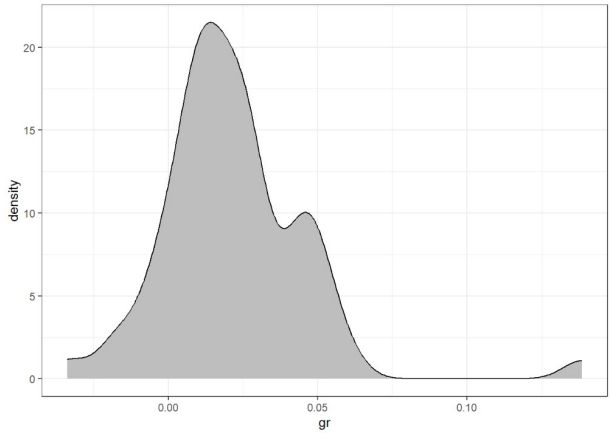


Figura 3.1: Distribuzione della variabile dipendente

3.3.1 Regressione OLS e multicollinearità

Inizialmente, tramite la funzione lm, viene stimata una regressione col metodo OLS che contiene tutte le variabili.

```
ols.model <- lm(gr ~ ., data = PWT)
summary(ols.model)
```

```
## Coefficients:
##                       Estimate         Std. Error t value Pr(>|t|)
## (Intercept) -13781.42712136626  43902.64554033106  -0.314  0.75791
## rgdpe18         -0.00000058293      0.00000020061  -2.906  0.01087 *
## rgdpe19          0.00000241731      0.00000286386   0.844  0.41189
## rgdpo19         -0.00000088509      0.00000074151  -1.194  0.25116
## pop19            0.00044274770      0.00034336281   1.289  0.21677
## emp19           -0.00123474053      0.00089399840  -1.381  0.18747
## avh19           -0.00002614379      0.00002093992  -1.249  0.23098
## hc19             0.01339344069      0.00647161860   2.070  0.05617 .
## ccon19           0.00000093069      0.00000091629   1.016  0.32587
## cda19           -0.00000145978      0.00000113315  -1.288  0.21718
## cgdpe19         -0.00000190671      0.00000278718  -0.684  0.50435
## cgdpo19          0.00000136951      0.00000084210   1.626  0.12470
## cn19             0.00000009132      0.00000007658   1.192  0.25160
## ck19             0.47547805731      0.46220838952   1.029  0.31992
## ctfp19           0.10073850907      0.19053647296   0.529  0.60474
## cwtfp19         -0.16040196479      0.20379359507  -0.787  0.44349
## rgdpna19        -0.00000031219      0.00000014586  -2.140  0.04916 *
## rconna19        -0.00000075749      0.00000083466  -0.908  0.37847
## rdana19          0.00000118510      0.00000100582   1.178  0.25705
## rnna19          -0.00000009009      0.00000007813  -1.153  0.26691
## rkna19           0.32069180715      0.10144998517   3.161  0.00646 **
## rtfpna19         0.13895407312      0.12819703095   1.084  0.29552
## rwtfpna19        0.09049235889      0.10520889344   0.860  0.40326
## labsh19          0.00429971544      0.04741289575   0.091  0.92894
## irr19            0.00695723896      0.14582919421   0.048  0.96258
## delta19          0.48739271006      0.34246048892   1.423  0.17514
## pl_con19         1.78648681133      0.55502649376   3.219  0.00574 **
## pl_da19         -0.66470230901      0.20633739198  -3.221  0.00571 **
## pl_gdpo19       -0.02109640663      0.06258418159  -0.337  0.74072
## csh_c19      13780.78886294513  43902.63327338610   0.314  0.75792
## csh_i19      13780.64036596236  43902.63390600061   0.314  0.75793
## csh_g19      13780.80324364954  43902.62469833600   0.314  0.75792
## csh_x19      13780.50413412859  43902.61060346731   0.314  0.75793
## csh_m19      13780.49027746033  43902.61042283056   0.314  0.75793
## csh_r19      13780.56891766399  43902.60644537343   0.314  0.75793
## pl_c19          -0.83155936007      0.27307249176  -3.045  0.00818 **
## pl_i19           0.13522287056      0.08365782416   1.616  0.12684
## pl_g19          -0.40291278288      0.13514869329  -2.981  0.00932 **
## pl_x19           0.13781544912      0.05543097481   2.486  0.02518 *
## pl_m19           0.04692143394      0.14448488024   0.325  0.74986
## pl_n19           0.02021215920      0.03935143546   0.514  0.61499
## pl_k19          -0.01472735793      0.01328792900  -1.108  0.28518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0069 on 15 degrees of freedom
## Multiple R-squared:  0.9795, Adjusted R-squared:  0.9236
## F-statistic:  17.5 on 41 and 15 DF,  p-value: 0.0000002173
```

Dal risultato ottenuto è possibile notare che, nonostante il valore dell'R quadro corretto e quello della statistica F siano elevati, molti coefficienti non sono statisticamente significativi. Questi

sono segnali della presenza di multicollinearità tra le variabili esplicative.
Tramite la creazione di una matrice di correlazione è possibile analizzare le relazioni che susst

```
subset <- PWT %>% select(rgdpe19, rgdpo19, pop19, emp19, rgdpe18, rgdpna19,
                         cgdpo19, delta19, csh_c19, pl_c19, pl_i19)
corr_matrix <- cor(subset, use="complete.obs")
corrplot(corr_matrix, method="circle", type="lower")
```

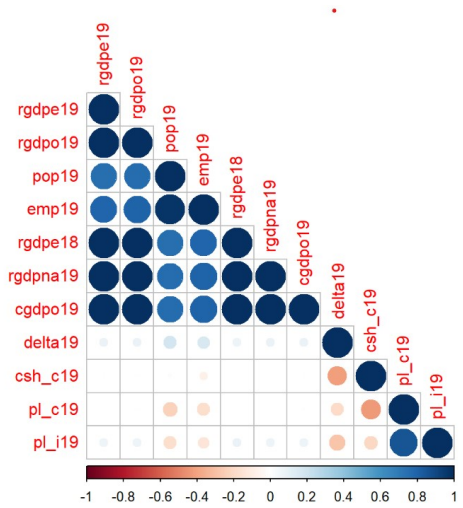


Figura 3.2: Questo grafico analizza la correlazione presente tra le variabili selezionate in base all'intensità del blu (correlazione positiva) e del rosso (correlazione negativa). Man mano che aumenta l'intensità del blu, aumenta la correlazione positiva tra le variabili considerate. Per esempio, la variabile $rgdpo19$ risulta essere altamente correlata con la variabile $cgdpo19$, mentre al contrario la variabile $delta19$ risulta essere debolmente correlata con la variabile $cgdpo19$.

Variance inflation factor

Un altro modo per scovare la multicollinerità è considerare la funzione VIF che calcola per ogni regressore la statistica VIF(The variance inflation factor) uguale a:

$$VIF(X_j) = \frac{1}{R^2_{X_j|X_{-j}}} \tag{3.3}$$

dove $R^2_{X_j|X_{-j}}$ è calcolato sulla regressione di una variabile esplicativa su tutte le altre. In caso di presenza di multicollinearità, il valore dell'R quadro sarà vicino a 1 e quindi il valore della statisica VIF sarà elevato. In particolare, valori che compresi tra il 5 e il 10 indicano la presenza di correlazione tra le variabili che potrebbe essere problematica, mentre i valori che superano il 10 indicano la presenza di multicollinearità.

```
vif(ols.model)
```

Dal seguente output della funzione VIF è possibile notare che molte variabili presentano un valore della statistica elevato, quindi non è opportuno inserire tutte le variabili all'interno del modello.

```
##      rgdpe18      rgdpe19      rgdpo19        pop19        emp19        avh19
## 7.259456e+05 1.654230e+08 1.077217e+07 9.737387e+03 1.425147e+04 2.989062e+01
##         hc19       ccon19        cda19      cgdpe19      cgdpo19         cn19
## 1.039431e+01 7.684102e+06 2.503203e+07 1.552069e+08 1.327265e+07 1.950945e+06
##         ck19       ctfp19      cwtfp19     rgdpna19     rconna19      rdana19
## 1.134100e+04 1.216509e+03 1.143626e+03 3.974190e+05 6.583591e+06 2.017495e+07
##       rnna19       rkna19     rtfpna19    rwtfpna19      labsh19        irr19
## 1.975219e+06 2.309240e+01 1.329520e+01 1.724449e+01 1.360646e+01 4.519031e+01
##      delta19     pl_con19      pl_da19    pl_gdpo19      csh_c19      csh_i19
## 1.240913e+01 2.907265e+04 3.124905e+03 3.126630e+02 2.489139e+13 1.222285e+13
##      csh_g19      csh_x19      csh_m19      csh_r19       pl_c19       pl_i19
## 5.890460e+12 3.303230e+14 3.479577e+14 2.193330e+13 5.966816e+03 3.171639e+02
##       pl_g19       pl_x19       pl_m19       pl_n19       pl_k19
## 2.772140e+03 6.540433e+00 7.058201e+01 1.103050e+02 1.315791e+01
```

3.3.2 Model selection

Prima di costruire qualsiasi modello, è necessario suddividere l'intero dataset in due sottocampioni, uno chiamato train dataset, che verrà utilizzato per la stima del modello, e un altro chiamato test dataset, che verrà utilizzato per la predizione. Il procedimento da seguire per la creazione e il confronto dei diversi modelli di regressione è il seguente:

1. Viene stimato un modello di regressione con il train dataset.
2. Vengono calcolati i valori predetti della variabile dipendente con suddetto modello, utilizzando il test dataset. In questo modo è possibile valutare la capacità predittiva del modello.
3. Con i valori predetti viene calcolato l'RMSE (Root Mean Squared Error) per permettere il confronto tra i modelli.

Dato che i dati sono disponibili in forma di cross-sectional, si utilizza la funzione createDataPartition che permettere di suddividere randomicamente train dataset e test dataset assegnando una percentuale p al train dataset e la rimanente al test dataset.

```
set.seed(19)
index <- createDataPartition(PWT$gr, p=0.8, list=FALSE, times=1)
traindata <- PWT[index,]
testdata <- PWT[-index,]
```

Per effettuare la selezione di variabili tramite il metodo di Subset Selection, viene utilizzato il pacchetto "leaps" che contiene la funzione regsubsets() che esegue forward, backward e stepwise selection, identificando il miglior modello che contiene un dato numero di variabili espicative. La sintassi è la seguente:

regsubsets(formula, nvmax, method)

Per formula si intende l'inserimento del modello con tutte le possibili variabili da includere, mentre nvmax esprime il massimo numero di variabili che si vogliono selezionare. Infine, con method è possibile specificare quale metodo di subset selection si vuole utilizzare per la selezione. In questo esempio verrà usato "seqrep" per la procedura ibrida Stepwise Selection.

```
best_model1 <- regsubsets(gr~., traindata, nvmax=41, method="seqrep")
model_summary <- summary(best_model1)
model_summary
```

L'output di model summary è rappresentato da una serie di righe nelle quali viene indicato tramite un asterisco se ogni variabile è inclusa nel modello o meno. Alla prima riga solo una variabile viene inclusa nel modello, successivamente nella seconda riga (quindi secondo modello) anche le variabili x e y vengono incluse. Il processo continua finchè tutte le variabili entrano a far parte del modello.
Per decidere quale modello è il migliore, è necessario definire un criterio, R^2 corretto o BIC, e considerare quello che performa meglio in base al criterio scelto.

Criterio R^2 corretto

Tramite la funzione *which.max*() è possibile ottenere il numero della riga di model summary che, tramite le variabili selezionate, permette di ottenere il massimo R^2 corretto, in questo caso la riga 31.

```
which.max(model_summary$adjr2)
plot(model_summary$adjr2, xlab="Numero delle variabili", ylab="R quadro corretto",
     type="o")
points(31, model_summary$adjr2[31], col="red")
```

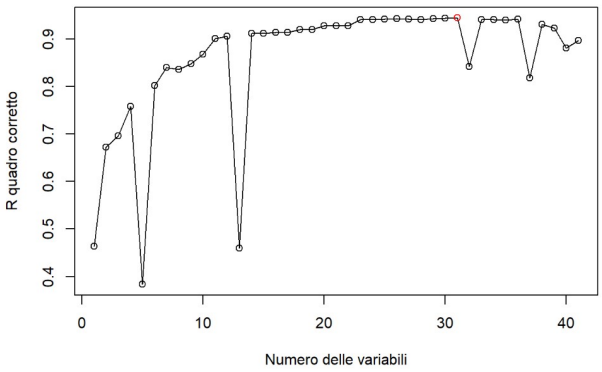


Figura 3.3: In questo grafico ogni pallino rappresenta l'R^2 corretto in funzione del numero di variabili che vengono inserite in ogni modello. Il pallino evidenziato in rosso rappresenta il punto di massimo dell'R^2 che si ottiene in corrispondenza di 31 variabili.

Infine, grazie alla funzione *coef*() vengono esplicitati i coefficienti della riga 31 e poi viene costruito il modello con i coefficienti selezionati.

```
coef(best_model1,31)
model.R <- lm(gr ~ rgdpe18+rgdpo19+emp19+hc19+ccon19+cda19+cgdpe19+cgdpo19+ctfp19+
              cwtfp19+rgdpna19+rconna19+rdana19+rnna19+rkna19+rwtfpna19+labsh19+delta19+
                pl_con19+pl_da19+csh_c19+csh_i19+csh_g19+csh_x19+
                csh_m19+csh_r19+pl_c19+pl_g19+pl_x19+pl_n19+pl_k19, data=traindata)
summary(model.R)
```

```
## Coefficients:
##                         Estimate          Std. Error t value    Pr(>|t|)
## (Intercept)  49678.686863087823  31011.669239585106   1.602    0.128726
## rgdpe18         -0.000000323356      0.000000105260  -3.072    0.007296 **
## rgdpo19         -0.000000512519      0.000000425356  -1.205    0.245755
## emp19           -0.000208594026      0.000071671124  -2.910    0.010217 *
## hc19             0.006741343102      0.003507239336   1.922    0.072582 .
## ccon19          -0.000000845100      0.000000528498  -1.599    0.129365
## cda19            0.000000698260      0.000000559834   1.247    0.230250
## cgdpe19          0.000000137045      0.000000105051   1.305    0.210494
## cgdpo19          0.000001078835      0.000000534574   2.018    0.060668 .
## ctfp19           0.164077618641      0.120840155053   1.358    0.193367
## cwtfp19         -0.210593118760      0.129377214450  -1.628    0.123108
## rgdpna19        -0.000000374558      0.000000115433  -3.245    0.005076 **
## rconna19         0.000000862591      0.000000497227   1.735    0.101994
## rdana19         -0.000000726762      0.000000529828  -1.372    0.189089
## rnna19           0.000000003746      0.000000001084   3.455    0.003257 **
## rkna19           0.519097246830      0.065468737259   7.929 0.000000622 ***
## rwtfpna19        0.151727662719      0.040651848520   3.732    0.001814 **
## labsh19          0.043990746372      0.026141304792   1.683    0.111821
## delta19          0.431482855316      0.253612796669   1.701    0.108227
## pl_con19         1.206414345248      0.262511753989   4.596    0.000298 ***
## pl_da19         -0.361418076716      0.101741391198  -3.552    0.002653 **
## csh_c19     -49679.485400650723  31011.683772456265  -1.602    0.128721
## csh_i19     -49679.690217965785  31011.679020837288  -1.602    0.128719
## csh_g19     -49679.440117144608  31011.680107968517  -1.602    0.128721
## csh_x19     -49679.759723968942  31011.656877087156  -1.602    0.128718
## csh_m19     -49679.781858147304  31011.656786948955  -1.602    0.128718
## csh_r19     -49679.716333852666  31011.654679891555  -1.602    0.128719
## pl_c19          -0.604431166164      0.150731037678  -4.010    0.001011 **
## pl_g19          -0.277663544635      0.066160211776  -4.197    0.000683 ***
## pl_x19           0.165830554816      0.036505034637   4.543    0.000333 ***
## pl_n19           0.053477062997      0.019107841041   2.799    0.012876 *
## pl_k19          -0.019932283155      0.009103096858  -2.190    0.043717 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.004583 on 16 degrees of freedom
## Multiple R-squared:  0.9811, Adjusted R-squared:  0.9445
## F-statistic: 26.78 on 31 and 16 DF,  p-value: 0.000000005076
```

Questo modello, nonostante riduca il numero di coefficienti, risulta essere ancora troppo complicato e non è in grado di soddisfare l'esigenza di parsimonia. Inoltre, il valore dell'R^2 è molto elevato ma molti coefficienti non sono statisticamente significativi, quindi è possibile concludere che il problema della multicollienarità sia ancora presente.

Criterio BIC

La funzione *which.min*() permette di ottenere il numero della riga di model summary che, tramite le variabili selezionate, restituisce il valore minimo di BIC, in questo caso la riga 11.

```
which.min(model_summary$bic)
plot(model_summary$bic, xlab="Numero di variabili", ylab="BIC", type="o")
points(11, model_summary$bic[11], col="red")
```

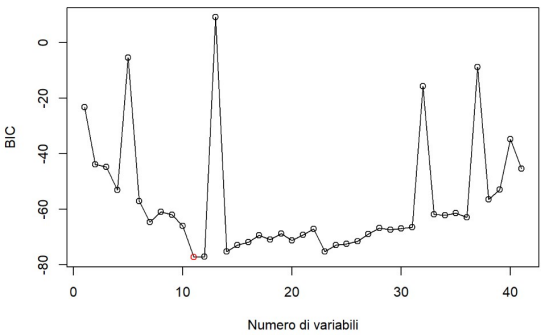


Figura 3.4: In questo grafico ogni pallino rappresenta il BIC in funzione del numero di variabili che vengono inserite in ogni modello. Il pallino evidenziato in rosso rappresenta il punto di minimo del bic che si ottiene in corrispondenza di 11 variabili

Infine vengono esplicitati i coefficienti della riga 11 e viene costruito il modello con i coefficienti selezionati.

```
coef(best_model1, 11)
modelbic <- lm(gr ~ rgdpe18+rgdpe19+cgdpe19+ck19+rkna19+rwtfpna19+csh_i19+
                 csh_m19+pl_x19+pl_m19+pl_n19, traindata)
summary(modelbic)
```

```
## Coefficients:
##                  Estimate    Std. Error t value        Pr(>|t|)
## (Intercept) -0.5060541864  0.0531186107  -9.527 0.0000000000223 ***
## rgdpe18     -0.0000003426  0.0000000574  -5.967 0.0000007686657 ***
## rgdpe19     -0.0000014353  0.0000002640  -5.436 0.0000039393832 ***
## cgdpe19      0.0000017571  0.0000003008   5.841 0.0000011352694 ***
## ck19         0.4045534445  0.0844430808   4.791 0.0000284196181 ***
## rkna19       0.3525699996  0.0433200046   8.139 0.0000000011204 ***
## rwtfpna19    0.1937985772  0.0315245458   6.148 0.0000004415852 ***
## csh_i19     -0.0898873317  0.0256728673  -3.501        0.001255 **
## csh_m19     -0.0108333982  0.0048727840  -2.223        0.032565 *
## pl_x19       0.1007452159  0.0233830049   4.308        0.000121 ***
## pl_m19      -0.1568500754  0.0275859144  -5.686 0.0000018275341 ***
## pl_n19       0.0078279862  0.0043125673   1.815        0.077838 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006122 on 36 degrees of freedom
## Multiple R-squared:  0.9241, Adjusted R-squared:  0.9009
## F-statistic: 39.84 on 11 and 36 DF,  p-value: < 0.00000000000000022
```

In questo modello l'R^2 aggiustato indica una bontà di adattamento molto buona e si assiste ad un grande miglioramento in termini di numerosità dei coefficienti i quali risultano essere statisticamente significativi.

3.3.3 Regressione ridge e Regressione lasso

Per realizzare un modello di regressione ridge o lasso, è necessario utilizzare la funzione glmnet che presenta la seguente sintassi:

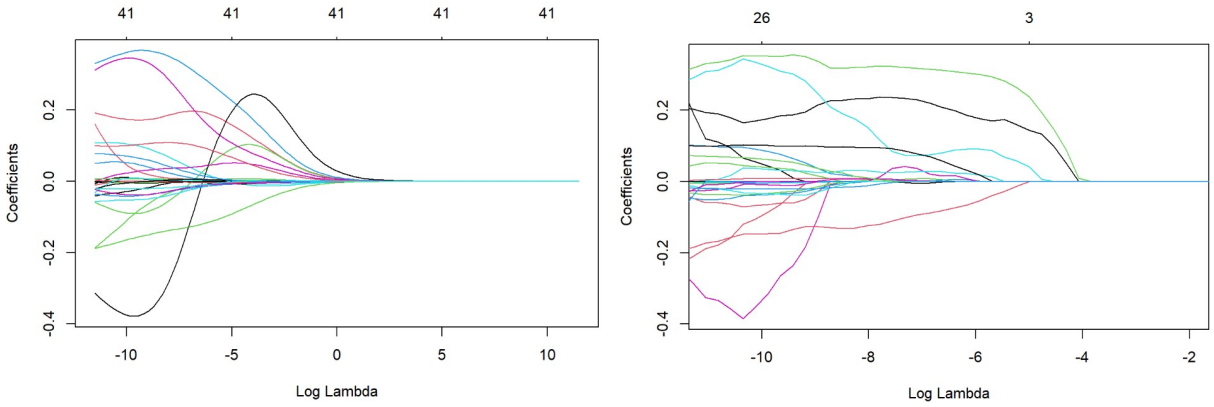
glmnet(x, y, alpha, lambda)

Dove:

- x rappresenta la matrice delle variabili esplicative per ogni unità statistica.
- y rappresenta il vettore della variabile dipendente per ogni unità statistica.
- Alpha è il parametro che definisce quale tecnica di regressione si vuole utilizzare. Per performare la regressione ridge $\alpha = 0$, per la regressione Lasso $\alpha = 1$ e un valore intermedio produce l'Elastic Net.
- Lambda rappresenta un vettore di valori di penalizzazione che viene definito precedentemente scegliendo l'intervallo e la numerosità.

Di seguito vengono creati gli input descritti precedentemente e i due modelli di regressione che vengono poi plottati per evidenziare l'andamento dei coefficienti delle due regressioni.

```
x <- model.matrix(gr ~., PWT)
y <- PWT$gr
grid <- 10^seq(5, -5, length=100)
ridge.mod <- glmnet(x, y, alpha=0, lambda=grid)
lasso.mod <- glmnet(x, y, alpha=1, lambda=grid)
```



Nel primo grafico viene plottato l'andamento dei coefficienti della regressione ridge in funzione di $log(\lambda)$. Man mano che il valore di $log(\lambda)$ aumenta, le stime dei coefficienti tendono a

zero senza mai raggiungerlo, infatti nell'asse superiore il numero di coefficienti rimane sempre uguale (41). Nel grafico a destra viene plottato l'andamento dei coefficienti della regressione lasso in funzione di $log(\lambda)$. Man mano che il valore di $log(\lambda)$ aumenta, le stime di alcuni coefficienti della regressione vengono stimati pari a zero, infatti nell'asse superiore il numero di coefficienti diminuisce man mano che $log(\lambda)$ aumenta.

Cross validation

La funzione trainControl, contenuta nel pacchetto caret, permette di specificare quale metodo di controllo si vuole utilizzare per l'addestramento dei dati. L'input *method* definisce il metodo di resampling da utilizzare, in questo caso Cross-validation. Tramite *number* viene indicato il numero di fold per la cross-validation e, infine, con *savePredictions* si indica se salvare o meno le predizioni del modello.

```
crossmethod <- trainControl(method = "cv", number=10, savePredictions = "all")
```

Successivamente, è necessario impostare la funzione train, parte del pacchetto caret, che permette di gestire in modo semplice e veloce l'addestramento, l'ottimizzazione e la validazione del modello, Kuhn and Johnson (2013). In questa funzione è necessario specificare:

- La formula del modello, quindi gr,che indica che si vuole prevedere la variabile growth rate utilizzando tutte le variabili presenti nel dataset.
- Il train dataset utilizzato per l'addestramento del modello.
- Il modello di regressione da utilizzare come glm per la regressione ridge, lasso o elastic net, lm per regressione lineare e rf per random forest.
- Il parametro tuneGrid serve creare una griglia di valori di alpha e di lambda da ottimizzare.
- trControl riprende il metodo di validazione incrociata definito precedentemente.

```
set.seed(1905)
ridge.model <- train(gr~ ., data=traindata, method="glmnet",
                     tuneGrid=expand.grid(alpha=0, lambda=grid),
                      trControl=crossmethod, na.action=na.omit)
#MIGLIOR VALORE DI LAMBDA CON ALPHA=0
ridge.model$bestTune
#PLOT RMSE IN FUNZIONE DI LOG-LAMBDA
plot(log(ridge.model$results$lambda), ridge.model$results$RMSE,
     xlab="log(lambda)", ylab="RMSE", xlim=c(-7,0))
```

Infine, è possibile ottenere il miglior parametro di tuning per la regressione Ridge e plottare un grafico che mostri l'RMSE in funzione di $log(lambda)$ per evidenziare che in corrispondenza del miglior parametro di tuning si ottiene l'RMSE minore.

	alpha <dbl>	lambda <dbl>
29	0	0.006734151

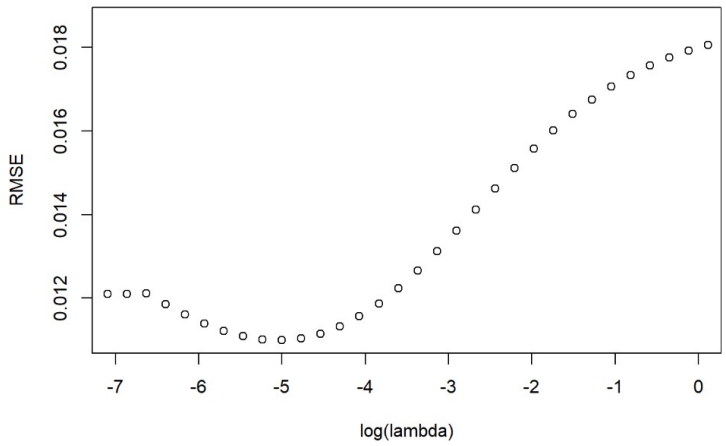


Figura 3.5: Questo grafico mostra l'RMSE in funzione del logaritmo del parametro di tuning. Il punto di minimo si ottiene in corrispondenza di lambda pari a 0.00, quindi per $log_e(\lambda) = -5$

Per esplicitare i coefficienti della regressione Ridge si utilizza la funzione coef() specificando il miglior valore di lambda e il modello model.ridge$finalModel, dove finalModel permette di selezionare il modello finale migliore tra tutti quelli che erano stati creati durante il processo di training.

```
round(coef(ridge.model$finalModel, ridge.model$bestTune$lambda), 7) #si applica il valore di lambda al modello costruito col
train dataset
```

```
## 42 x 1 sparse Matrix of class "dgCMatrix"
##                      s1
## (Intercept) -0.4903232
## rgdpe18      0.0000000
## rgdpe19      0.0000000
## rgdpo19      0.0000000
## pop19        0.0000008
## emp19       -0.0000001
## avh19        0.0000019
## hc19        -0.0014203
## ccon19       0.0000000
## cda19        0.0000000
## cgdpe19      0.0000000
## cgdpo19      0.0000000
## cn19         0.0000000
## ck19        -0.0008343
## ctfp19      -0.0036793
## cwtfp19      0.0038113
## rgdpna19     0.0000000
## rconna19     0.0000000
## rdana19      0.0000000
## rnna19       0.0000000
## rkna19       0.1440210
## rtfpna19     0.1596495
## rwtfpna19    0.1527929
## labsh19      0.0091533
## irr19        0.0739365
## delta19      0.0036446
## pl_con19    -0.0008548
## pl_da19     -0.0004868
## pl_gdpo19    0.0002824
## csh_c19      0.0283234
## csh_i19      0.0143397
## csh_g19      0.0089937
## csh_x19      0.0017392
## csh_m19     -0.0050820
## csh_r19     -0.0007692
## pl_c19      -0.0012230
## pl_i19      -0.0024909
## pl_g19      -0.0005312
## pl_x19       0.0662791
## pl_m19      -0.0571786
## pl_n19       0.0041537
## pl_k19       0.0021558
```

Dall'output emerge che molti coefficienti tendono verso lo zero. Tuttavia, non si osserva una vera e propria selezione delle variabili, poiché quelle con coefficienti estremamente bassi rimangono comunque incluse nel modello.

Regressione Lasso

Utilizzando il medesimo procedimento descritto precedentemente, è possibile costruire la regressione Lasso, impostando il parametro $\alpha = 1$.

```
set.seed(19)
lasso.model <- train(gr~ ., data=traindata, method="glmnet",
            tuneGrid=expand.grid(alpha=1,lambda=grid), trControl=crossmethod,
            na.action=na.omit)
#MIGLIOR VALORE DI LAMBDA CON ALPHA=1
lasso.model$bestTune
#PLOT RMSE IN FUNZIONE DI LOG-LAMBDA
plot(log(lasso.model$results$lambda), lasso.model$results$RMSE, xlab="log(lambda)", ylab="RMSE", xlim=c(-9,-4))
```

	alpha <dbl>	lambda <dbl>
19	1	0.0006579332

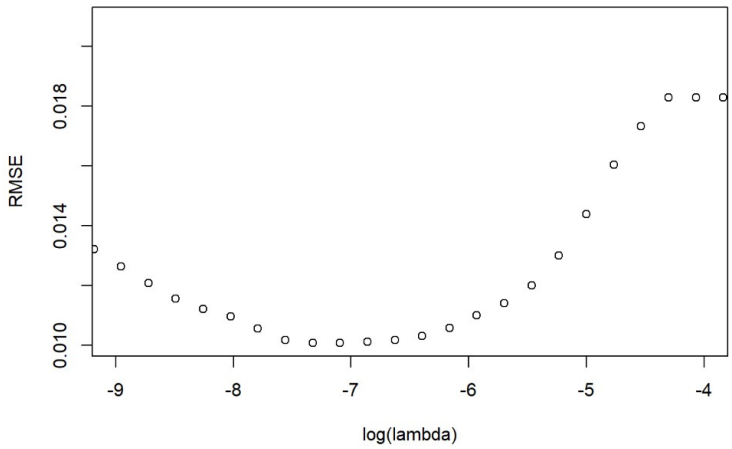


Figura 3.6: Questo grafico mostra l'RMSE in funzione di $log(\lambda)$. Il punto di minimo si ottiene in corrispondenza di lambda pari a 0.000, quindi per $log_e(\lambda) = -7.33$

Esplicitando i coefficienti della regressione è possibile notare che, a differenza della regressione ridge, molti sono stimati esattamente pari zero, quindi viene effettuata una selezione di variabili.

```
round(coef(lasso.model$finalModel, lasso.model$bestTune$lambda), 4) #si applica il valore di lambda al modello costruito col
train dataset
```

```
## 42 x 1 sparse Matrix of class "dgCMatrix"
##                    s1
## (Intercept) -0.5746
## rgdpe18      .
## rgdpe19      .
## rgdpo19      .
## pop19        .
## emp19        .
## avh19        .
## hc19         .
## ccon19       .
## cda19        .
## cgdpe19      .
## cgdpo19      .
## cn19         .
## ck19         .
## ctfp19       .
## cwtfp19      .
## rgdpna19     .
## rconna19     .
## rdana19      .
## rnna19       .
## rkna19       0.2071
## rtfpna19     0.1440
## rwtfpna19    0.1957
## labsh19      .
## irr19        0.0494
## delta19      .
## pl_con19     .
## pl_da19      .
## pl_gdpo19    .
## csh_c19      0.0316
## csh_i19      .
## csh_g19      .
## csh_x19      .
## csh_m19     -0.0041
## csh_r19      .
## pl_c19       .
## pl_i19       .
## pl_g19       .
## pl_x19       0.0855
## pl_m19      -0.0775
## pl_n19       0.0023
## pl_k19       .
```

Dato che le stime dei coefficienti della regressione Lasso sono distorte, viene utilizzata la regressione Post-Lasso che combina la Regressione Lasso e la Regressione OLS per eliminare il bias. Alcuni dei coefficienti della regressione lasso non sono stati inclusi nel modello in quanto non statisticamente significativi.

```
# Esecuzione della regressione lineare sulle variabili selezionate dalla Lasso
postlasso.model <- lm(gr ~ rkna19 +rwtfpna19+ pl_x19 + pl_m19 + csh_c19, traindata)
summary(postlasso.model)
```

```
## Coefficients:
##             Estimate Std. Error t value        Pr(>|t|)
## (Intercept) -0.54142    0.05676  -9.539 0.0000000000045 ***
## rkna19       0.23212    0.03729   6.225 0.0000001885925 ***
## rwtfpna19    0.28043    0.03677   7.627 0.0000000018674 ***
## pl_x19       0.12555    0.03001   4.183        0.000143 ***
## pl_m19      -0.11505    0.02566  -4.483 0.0000559693437 ***
## csh_c19      0.03938    0.01663   2.368        0.022585 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008638 on 42 degrees of freedom
## Multiple R-squared:  0.8236, Adjusted R-squared:  0.8027
## F-statistic: 39.23 on 5 and 42 DF,  p-value: 0.000000000000008937
```

Infine, l'RMSE viene impiegato come criterio per il confronto tra i diversi modelli sviluppati. Di conseguenza, è necessario calcolare i valori predetti per ciascun modello e confrontarli con i valori osservati. Il dataset utilizzato è il testdata: questo permette di capire se il modello costruito con i dati di addestramento è accurato anche per i nuovi dati.

```
predictions1 <- predict(ridge.model, newdata=testdata)
predictions2 <- predict(postlasso.model, newdata=testdata)
predictions3 <- predict(model.R, newdata=testdata)
predictions4 <- predict(modelbic, newdata=testdata)
model_accurancy <- data.frame(RMSE_ridge=RMSE(predictions1, testdata$gr),
                  RMSE_postlasso=RMSE(predictions2, testdata$gr),
                  RMSE_modelR=RMSE(predictions3, testdata$gr),
                  RMSE_bic=RMSE(predictions4, testdata$gr))
model_accurancy
```

RMSE_ridge <dbl>	RMSE_postlasso <dbl>	RMSE_modelR <dbl>	RMSE_bic <dbl>
0.03361692	0.02654154	0.01875356	0.0227704

Dalla tabella è possibile notare che, escludendo modelR per i problemi legati alla multicollinearità, gli RMSE minori si ottengono dal modello di regressione post lasso e dal modello di regressione che utilizza il BIC come criterio di model selection.

3.3.4 Interpretazione economica dei coefficienti della Regressione Post Lasso

Le variabili inserite nel modello post lasso sono le seguenti:

- Rkna (Revised Capital National Accounts) è una variabile che misura la quantità di investimenti in capitale fisico e il loro tasso di deprezzamento. Questa variabile ha un effetto positivo di 0.23212 sul tasso di crescita del pil.

- Rwtfpna (Welfare-relevant Total Factor Productivity at constant National Accounts prices) misura l'impatto della produttività totale dei fattori considerando anche il benessere sociale della collettività. Questa variabile ha un effetto positivo di 0.28043 sul tasso di crescita del pil.
- Pl_x (Price level of exports) è una variabile che definisce il prezzo delle esportazioni e influisce positivamente sul tasso di crescita del pil con un coefficiente pari a 0.12555.
- Pl_m (Price level of imports) rappresenta il prezzo delle importazioni e ha un impatto negativo sul tasso di crescita del PIL pari a -0.11505.
- Csh_c misura la quota di consumo delle famiglie rispetto al PIL. Questa variabile ha un impatto positivo pari a 0.03938 sul tasso di crescita del pil.

Analizzare queste variabili considerando la teoria macroeconomica permette di comprendere perchè la regressione lasso ha selezionato queste variabili e perchè sono rilevanti al fine della predizione del tasso di crescita del pil.

Il consumo

La variabile csh_c misura la quota di consumo e rappresenta uno dei componenti principali del PIL, arrivando a rappresentarne il 60-70%. Il fattore principale da cui dipende il consumo delle famiglie è il reddito disponibile dato dalla differenza tra il reddito aggregato e le imposte al netto dei trasferimenti. E' possibile sintetizzare la relazione tra il consumo e il reddito disponibile con la seguente funzione:

$$C = c_0 + c_1 Y_d \tag{3.4}$$

Dove C è il consumo e Y_d è il reddito disponibile che influenza positivamente il consumo. I parametri c_0 e c_1 rappresentano rispettivamente la componente autonoma del consumo e la propensione marginale al consumo, ovvero la variazione del consumo a fronte della variazione unitaria del reddito disponibile. Il parametro c_1 è compreso tra zero e uno perchè un aumento del reddito disponibile genera un aumento del consumo, tuttavia tale aumento è meno che proporzionale.

Perchè il consumo ha un impatto significativo sul PIL? A fronte di un aumento del consumo, le imprese aumentano la produzione che a sua volta porta ad un aumento del reddito. L'aumento del reddito genera un secondo aumento del consumo pari all'aumento iniziale moltiplicato per la propensione marginale al consumo. Il processo continua con un secondo aumento della produzione e quindi del reddito, producendo un effetto a catena noto come effetto moltiplicatore. E' possibile concludere che il moltiplicatore è in grado di amplificare l'effetto dell'aumento iniziale in una componente della domanda aggregata sul livello del PIL.
L'effetto moltiplicatore può essere espresso matematicamente come:

$$\frac{1}{1-c_1} \tag{3.5}$$

Quanto è maggiore la propensione marginale al consumo, tanto sarà maggiore l'effetto del moltiplicatore sull'economia.

Relazione tra produzione e investimenti

Rkna è una variabile che misura la quantità di investimenti in beni capitali e il deprezzamento del capitale esistente. Gli investimenti dipendono principalmente da due fattori: il livello delle vendite e il tasso di interesse. Un aumento delle vendite può portare le imprese ad ampliare la loro capacità produttiva per sostenere la domanda crescente, investendo in nuovi macchinari, impianti o tecnologie. Il tasso di interesse, ovvero il costo degli investimenti, è un elemento fondamentale: tassi di interesse elevati rendono gli investimenti meno attraenti, mentre tassi di interesse più bassi possono incentivare gli investimenti. Per riflettere questi due effetti, la relazione di investimento è formulata come segue:

$$I = I(\underset{+}{Y}, \underset{-}{i}) \tag{3.6}$$

L'equazione afferma che l'investimento dipende positivamente dalla produzione Y e negativamente dal tasso di interesse i. Avendo introdotto l'investimento, è possibile esprimere il PIL come:

$$Y = C(Y - T) + I(Y, i) + \bar{G} \tag{3.7}$$

L'equazione rappresenta la relazione IS (Investment-Savings) ampliata che mostra tutte le combinazioni di tasso di interesse e livello di produzione per cui il mercato dei beni è in equilibrio. La variabile $\bar{G}$ rappresenta la spesa pubblica che non è all'interno del modello e quindi viene considerata come una variabile esogena.

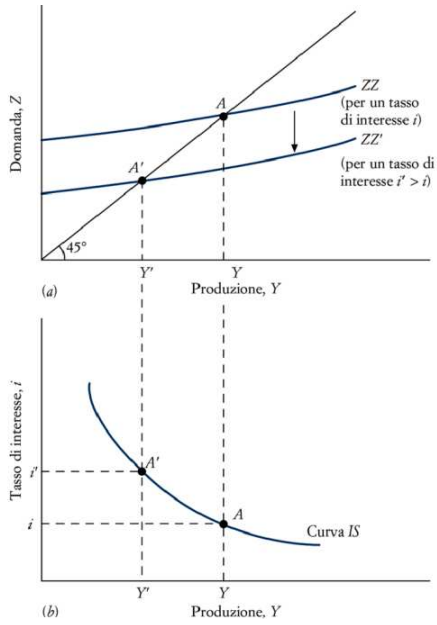


Figura 3.7: Se il tasso di interesse aumenta, gli investimenti diminuiscono con un impatto negativo sulla domanda aggregata e sul reddito, Blanchard and Giavazzi (2021).

Il livello degli investimenti influenza il tasso di crescita anche nel lungo periodo attraverso l'accumulazione del capitale. Assumendo una funzione di investimento proporzionale alla produzione $I_t = sY_t$, dove s è il tasso di risparmio, è possibile definire la relazione tra investimento e accumulazione del capitale come segue:

$$K_{t+1} = (1-\delta)K_t + I_t \tag{3.8}$$

Solow (1956). Il livello di capitale al tempo t+1 è dato dal livello di capitale al tempo t al netto del tasso di deprezzamento più il livello degli investimenti al tempo t.
Per ottenere la relazione tra produzione e accumulazione del capitale e vedere come il livello degli investimenti influenza la produzione è necessario definire la relazione tra la produzione e capitale per lavoratore al tempo t:

$$\frac{Y_t}{N} = f(\frac{K_t}{N}) \tag{3.9}$$

dove Y_t è la produzione al tempo t, K è il capitale e N è l'occupazione. Secondo questa funzione è il capitale a determinare l'ammontare della produzione nel lungo periodo.
La relazione tra produzione e accumulazione del capitale si ottiene combinando le equazioni 3.8 e 3.9 e riordinando i termini si ottiene:

$$\frac{K_{t+1}}{N} - \frac{K_t}{N} = s\frac{Y_t}{N} - \delta\frac{K_t}{N} \quad \rightarrow \quad \frac{K_{t+1}}{N} - \frac{K_t}{N} = sf(\frac{K_t}{N}) - \delta\frac{K_t}{N} \tag{3.10}$$

La variazione dello stock di capitale è data dall'investimento per lavoratore al netto del deprezzamento del capitale per lavoratore.
Se l'investimento per lavoratore è maggiore del deprezzamento per lavoratore allora la variazione del capitale per lavoratore è positiva portando ad un aumento dello stock di capitale.
Il tasso di risparmio s influenza il livello della produzione per lavoratore, ma non esercita alcun effetto sul tasso di crescita di lungo periodo.
E' bene precisare che il prodotto per lavoratore non dipende solo dal capitale fisico (K) ma anche dal capitale umano (H). Inoltre, l'accumulazione del capitale permette di aumentare la produzione, ma non è sufficiente per sostenere la crescita, per la quale è imprescindibile il progresso tecnologico.

La produttività totale dei fattori

La variabile rwtfpna considera l'impatto della produttività totale dei fattori. Come detto precedentemente, la produttività non può essere spiegata solamente in termini di capitale e lavoro, ma è necessario considerare anche il progresso tecnologico che consente di generare una quantità maggiore di produzione a parità di fattori produttivi, permette di aumentare la qualità della produzione e di creare nuovi prodotti o ampliare la gamma di prodotti già esistenti. Nonostante l'analisi dell'impatto della tecnologia sulla produttività sia tipicamente un'analisi di lungo periodo, secondo Prescott (1986) le fluttuazioni a breve termine del tasso di crescita possono essere spiegate anche da shock tecnologici esogeni, cioè improvvisi cambiamenti nella produtti-

vità tecnologica. Per esempio, l'adozione di tecnologie dell'informazione e delle comunicazioni (ICT) a partire dagli anni 90 ha determinato in pochi anni un aumento significativo della produttività in molti settori.
Secondo il modello di Solow con progresso tecnologico, Solow (1956), la funzione di produzione viene riscritta come:

$$Y_t = f(K_t, A_t N_t) \tag{3.11}$$

dove A_t è lo stato della tecnologia o la produttività al tempo t e il termine $A_t N_t$ indica il lavoro effettivo al tempo t. In questo modello vengono introdotti due nuovi elementi: il tasso di crescita della tecnologia g_A e il tasso di crescita della popolazione g_N.
Per integrare gli indicatori di benessere e definire la variabile inclusa in questo modello di regressione, è necessario introdurre una funzione di benessere sociale data da:

$$W_t = f(u_{1t}, u_{2t}, ..., u_{nt}) \tag{3.12}$$

dove W_t è il benessere sociale al tempo t e u_i con $i = (1, 2.., n)$ è la funzione di utilità dei singoli individui, Bergson (1938). Data la funzione di benessere sociale, la produttività totale dei fattori può essere definita da una funzione Cobb-Douglas:

$$Y_t = K_t^{\alpha} * (A_t N_t)^{\beta} * W_t^{\gamma} \tag{3.13}$$

Dove $Y_t, K_t, A_t N_t$ e W_t sono rispettivamente la produzione, il capitale, il lavoro effettivo e il benessere sociale al tempo t, mentre α, β e γ sono i pesi associati a ciascuna variabile, Van Beveren (2012).
A lungo termine l'adeguamento al progresso tecnologico avviene attraverso un aumento della produzione. Nel breve termine il progresso tecnologico ha effetti sia sulla produzione sia sull'occupazione, Blanchard and Johnson (2013). L'impatto di un cambiamento nel livello della tecnologia nel breve periodo può essere analizzato assumendo che la produzione sia funzione del lavoro (N) e del progresso tecnologico (A) senza considerare il livello di capitale: $Y = AN$. Riscrivendo l'equazione come $N = Y/A$ è possibile esprimere l'occupazione come rapporto tra il livello di produzione e della produttività. A parità di produzione, un incremento della produttività comporta una riduzione del livello di occupazione. Attraverso il seguente modello è possibile analizzare come (nel breve periodo) un aumento della produttività influisce sulla produzione e sull'occupazione.

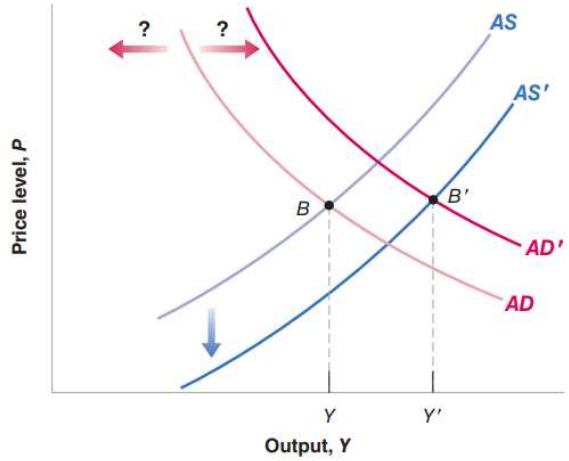


Figura 3.8: La curva AD rappresenta la domanda aggregata, mentre la curva AS rappresenta l'offerta aggregata. L'effetto di un aumento della produttività sull'offerta aggregata è quello di ridurre la quantità di lavoro necessario per produrre un'unità di produzione, riducendo i costi per le imprese. Ciò induce le imprese a ridurre il prezzo che esse applicano a qualsiasi livello di produzione. Di conseguenza, la curva dell'offerta aggregata si sposta verso il basso, da AS a AS'. Un incremento della produttività non esercita un effetto univoco sulla domanda aggregata, poiché il suo impatto dipende dalla causa sottostante a tale incremento. Se la crescita della produttività è attribuibile a un'innovazione tecnologica, è probabile che si assisterà a un'espansione della domanda, poiché le aspettative di una maggiore crescita futura tendono a rafforzare l'ottimismo dei consumatori. Inoltre, la prospettiva di profitti futuri più elevati può stimolare un'accelerazione degli investimenti. Se l'incremento della produttività è dovuto a un uso più efficiente delle tecnologie già esistenti, è probabile che si verifichi una contrazione della domanda aggregata. La riorganizzazione dei processi produttivi potrebbe determinare una riduzione dell'occupazione, con conseguente diminuzione dei redditi disponibili e della propensione al consumo. In tale scenario, la curva della domanda aggregata potrebbe spostarsi verso sinistra.

L'evidenza empirica: Lo studio Brynjolfsson and Hitt (2003) si focalizza sull'impatto che ha l'adozione delle tecnologie dell'informazione e della comunicazione (ICT) nella produttività e nella crescita della produzione, utilizzando dati di 527 aziende statunitensi. Il risultato principale evidenzia che l'utilizzo di ICT contribuisce alla crescita della produzione a breve termine, tuttavia tali effetti dipendono anche dalla modalità di integrazione e gestione delle tecnologie all'interno dell'organizzazione. È quindi necessario che le imprese non solo investano in tecnologie avanzate, ma adattino anche i loro processi e strutture organizzative per massimizzare i benefici derivanti dalle nuove tecnologie.

Prezzo delle esportazioni e delle importazioni

L'apertura del mercato a livello internazionale determina un cambiamento nell'equilibrio economico, in quanto influisce sulla scelta dei consumatori che devono decidere se acquistare beni nazionali o esteri. Un aumento del livello dei prezzi delle importazioni incide negativamente sul tasso di crescita del PIL. Per esempio, l'aumento del costo delle materie prime importate provoca un aumento dei costi di produzione e quindi una riduzione dell'utile che si può tradurre una diminuzione degli investimenti. Inoltre, un aumento del prezzo dei beni importati diminuisce il

potere d'acquisto delle famiglie, influenzando negativamente la componente dei consumi. Per quanto riguarda le esportazioni, un aumento del livello dei prezzi delle esportazioni garantisce alle imprese esportatrici maggiori ricavi e quindi maggiori profitti. Il possibile aumento degli investimenti che ne consegue stimola l'attività economica, la competitività e la crescita del PIL. Importazioni ed esportazioni possono essere definite come segue:

$$IM = IM(Y, \varepsilon), \qquad X = X(Y^*, \varepsilon) \tag{3.14}$$

Le importazioni sono definite da una funzione del reddito nazionale (Y) e del tasso di cambio reale (ε). Per quanto riguarda le esportazioni, si considera il reddito estero (Y*) e il tasso di cambio reale (ε). E' evidente che la variabile che influenza sia importazioni che esportazioni è il tasso di cambio reale, tuttavia l'impatto è differente.
Il tasso di cambio reale è "il prezzo dei beni nazionali in termini di beni esteri" (Blanchard and Giavazzi, 2021) che si differenzia dal tasso di cambio nominale, ovvero "il prezzo della moneta nazionale espresso in termini di moneta estera". Il tasso di cambio reale ε viene calcolato moltiplicando il tasso di cambio nominale (E) per il livello dei prezzi nazionali (P) rapportato al livello dei prezzi esteri (P*):

$$\varepsilon = E\frac{P}{P*} \tag{3.15}$$

Le variazioni del tasso di cambio reale possono essere in aumento (apprezzamento reale) o in diminuzione (deprezzamento reale). Un apprezzamento reale rende i beni esteri più economici e i beni domestici più costosi, mentre un deprezzamento reale rende i beni esteri più costosi rispetto ai beni domestici.
Un apprezzamento del tasso di cambio reale influenza positivamente le importazioni e negativamente le esportazioni, mentre un deprezzamento del tasso di cambio reale genera un aumento delle esportazioni e una diminuzione delle importazioni. E' possibile riassumere la variazione di esportazioni ed importazioni considerando le esportazioni nette o bilancia commerciale:

$$NX = X(\underset{+}{Y^*}, \underset{-}{\varepsilon}) - \frac{IM(\underset{+}{Y}, \underset{+}{\varepsilon})}{\varepsilon} = NX(\underset{+}{Y^*}, \underset{+}{Y}, \underset{-}{\varepsilon}) \tag{3.16}$$

Un deprezzamento reale provoca un aumento delle esportazioni, una diminuzione delle importazioni e un aumento del prezzo relativo dei beni esteri in termini di beni nazionali ($\frac{1}{\varepsilon}$) che provoca un aumento del valore delle importazioni. Si può concludere che un deprezzamento del tasso di cambio reale genera un aumento delle esportazioni nette[1].
Il tasso di cambio e il reddito (nazionale ed estero) non sono gli unici fattori in grado di influenzare la bilancia commerciale. Per un'analisi completa, è necessario considerare anche le politiche commerciali, le politiche fiscali e monetarie e il costo dei trasporti e della logistica.

[1] secondo la condizione di Marshall-Lerner

3.4 Conclusioni

Nella presente tesi sono stati esaminati i concetti fondamentali dei Big Data e le tecniche di regressione più adatte per la loro analisi. È stato dimostrato che i Big Data costituiscono una risorsa sempre più importante per la realizzazione di analisi dettagliate a supporto del processo decisionale, in tutti i settori. Tuttavia, sono state evidenziate anche le sfide e le problematiche associate a tale ambito. I dataset ad alta dimensionalità, quando gestiti con le tecniche di regressione tradizionali, presentano problemi di overfitting e difficoltà di interpretazione. Pertanto, sono state analizzate tecniche avanzate di selezione delle variabili e di Shrinkage.
Il confronto tra la regressione ridge e la regressione lasso ha messo in luce come quest'ultima, a differenza della prima, sia in grado di selezionare le variabili più rilevanti per la predizione. La regressione lasso è stata approfondita attraverso l'introduzione della regressione post-lasso, per affrontare il problema della distorsione delle stime dei coefficienti, e della regressione fused lasso, per la gestione delle serie temporali. Il secondo capitolo conclude con una discussione sulle tecniche di machine learning, come gli alberi decisionali che rappresentano un ulteriore progresso e consentono l'apprendimento supervisionato.
Nell'ultimo capitolo, alcune delle tecniche di regressione precedentemente discusse sono state applicate al dataset Penn World Table per la previsione del tasso di crescita del PIL a breve termine. I risultati ottenuti hanno dimostrato che la selezione delle variabili tramite la regressione post-lasso può significativamente migliorare la capacità predittiva rispetto ai metodi tradizionali. Le variabili selezionate sono state analizzate in base alla teoria macroeconomica, che ha confermato la loro influenza sulla variabile dipendente. Oltre a consumo, investimenti ed esportazioni nette, è emerso come l'impatto della produttività totale dei fattori, che include il progresso tecnologico e indicatori di benessere collettivo, sia rilevante.
La presente ricerca si è focalizzata sull'analisi della crescita a breve termine e sull'utilizzo delle tecniche di regolarizzazione, tuttavia l'accuratezza del modello potrebbe essere ulteriormente migliorata impiegando un modello di machine learning, come il gradient boosting, in grado di gestire un dataset più complesso che consideri per ogni paese una serie temporale di dati (panel dataset).

[1] Numero parole utilizzate: 9938

Tabella variabili esplicative

Nome variabile	Descrizione	Media	DeviazioneStandard
rgdpe18	Expenditure-side real GDP at chained PPPs (in mil. 2017USD) in 2018	1843638.005	3815186.599
rgdpe19	Expenditure-side real GDP at chained PPPs (in mil. 2017USD) in 2019	1880838.565	3913563.817
rgdpo18	Output-side real GDP at chained PPPs (in mil. 2017USD)	1871176.251	3916065.187
pop18	Population (in millions)	91.217	260.221
emp18	Number of persons engaged (in millions)	42.408	123.931
avh18	Average annual hours worked by persons engaged	1811.431	240.810
hc18	Human capital index, based on years of schooling and returns to education	3.227	0.435
ccon18	Real consumption of households and government, at current PPPs (in mil. 2017USD)	1334553.720	2722015.347
cda18	Real domestic absorption, (real consumption plus investment), at current PPPs (in mil. 2017USD)	1856094.220	3932284.398
cgdpe18	Expenditure-side real GDP at current PPPs (in mil. 2017USD)	1873736.849	3894429.378
cgdpo18	Output-side real GDP at current PPPs (in mil. 2017USD)	1866748.760	3900385.833
cn18	Capital stock at current PPPs (in mil. 2017USD)	8181819.112	16501666.529
ck18	Capital services levels at current PPPs (USA 1)	0.101	0.207
ctfp18	TFP level at current PPPs (USA=1)	0.737	0.168
cwtf18	Welfare-relevant TFP levels at current PPPs (USA=1)	0.704	0.152
rgdpna18	Real GDP at constant 2017 national prices (in mil. 2017USD)	1876861.797	3939727.802
rconna18	Real consumption at constant 2017 national prices (in mil. 2017USD)	1363130.909	2833767.549
rdana18	Real domestic absorption at constant 2017 national prices (in mil. 2017USD)	1892934.364	4099379.920
rnna18	Capital stock at constant 2017 national prices (in mil. 2017US)	8187484.128	16517138.038
rkna18	Capital services at constant 2017 national prices (2017=1)	1.065	0.044
rtfpna18	TFP at constant national prices (2017=1)	1.003	0.026
rwtfpna18	Welfare-relevant TFP at constant national prices (2017=1)	1.005	0.036
labsh18	Share of labour compensation in GDP at current national prices	0.549	0.072
irr18	Real internal rate of return	0.098	0.042
delta18	Average depreciation rate of the capital stock	0.045	0.009
pl_con18	Price level of CCON (PPP/XR), price level of USA GDPo in 2017=1	0.747	0.281
pl_da18	Price level of CDA (PPP/XR), price level of USA GDPo in 2017=1	0.722	0.248
pl_gdpo18	Price level of CGDPo (PPP/XR), price level of USA GDPo in 2017=1	0.748	0.258
csh_c18	Share of household consumption at current PPPs	0.557	0.105
csh_i18	Share of gross capital formation at current PPPs	0.251	0.074
csh_g18	Share of government consumption at current PPPs	0.190	0.051
csh_x18	Share of merchandise exports at current PPPs	0.392	0.351
csh_m18	Share of merchandise imports at current PPPs	-0.441	0.359
csh_r18	Share of residual trade and GDP statistical discrepancy at current PPPs	0.050	0.098
pl_c18	Price level of household consumption, price level of USA GDPo in 2017=1	0.752	0.259
pl_i18	Price level of capital formation, price level of USA GDPo in 2017=1	0.683	0.196
pl_g18	Price level of government consumption, price level of USA GDPo in 2017=1	0.744	0.356
pl_x18	Price level of exports, price level of USA GDPo in 2017=1	0.669	0.042
pl_m18	Price level of imports, price level of USA GDPo in 2017=1	0.632	0.054
pl_n18	Price level of the capital stock, price level of USA in 2017=1	0.565	0.246
pl_k18	Price level of the capital services, price level of USA=1	0.736	0.252

Bibliografia

AIFIRM. Big data advanced analytics per il risk management. Technical report, AIFIRM, 2022.

Valentina Aprigliano, Guerino Ardizzi, and Libero Monteforte. Using the payment system data to forecast the economic activity. *International Journal of Central Banking*, 12(4):111–147, 2016.

Belloni and Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli 2013, Vol. 19, No. 2, 521-547*, 2013.

A. Bergson. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics, 52(2), 310-334*, 1938.

Amighini Blanchard and Giavazzi. *Scoprire la macroeconomia*. Il Mulino, 2021.

Oliver Blanchard and David R. Johnson. *Macroeconomics*. Pearson, 2013.

Brynjolfsson and Hitt. Computing productivity: Firm-level evidence. *The Review of Economics and Statistics*, 2003.

Capgemini. World payments report. Technical report, Capgemini, 2023.

Choi and Varian. Predicting the present with Google Trends. *Economic Record*, 88(s1):2–9, 2012.

Pierre Deville, Catherine Linard, Sylvie Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

EURONA. Eurostat review on national accounts and macroeconomic indicators. Technical report, European Union, 2017.

T. Hastie G. James, D. Witten and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynn Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457 (7232):1012–1014, 2009.

William H. Greene. *Econometric Analysis*. Pearson Education, 2002.

McArdle Kitchin. What makes big data, big data? exploring the ontological characteristics of 26 datasets. *Big Data Society, 3(1)*, 2016.

Max Kuhn and Kjell Johnson. *Applied Predictive Modelling*. Springer, 2013.

Alexandre L'Heureux, Mark Scholl, and Thierry de Boissieu. Machine learning with big data: Challenges and approaches. *IEEE Access*, 2017.

Edward Prescott. Theory ahead of business cycle measurement. *Quarterly Review*, 1986.

Rydning Reinsel, Gantz. Data age 2025:the evolution of data to life-critical. Technical report, IDC, 2017.

Robert Inklaar Robert C. Feenstra and Marcel P. Timmer. The next generation of the penn world table. *American Economic Review*, 2015.

Torsten Schmidt and Simeon Vosen. Forecasting private consumption: Survey-based indicators vs. google trends. *Journal of Forecasting*, 30(6):565–578, 2011.

Robert M. Solow. A contribution to the theory of economic growth. *The Quarterly Journal of Economics*, 70(1):65–94, 1956.

N. K. Suryadevara, S. C. Mukhopadhyay, S. D. T. Kelly, and S. P. S. Gill. Wsn-based smart sensors and actuator for power management in intelligent buildings. *IEEE/ASME Transactions on Mechatronics*, 20(2):564–571, 2013.

R. Tibshirani T. Hastie and J. Friedman. *The Elements of Statistical Learning*. Springer New York Inc, 2001.

I. Van Beveren. Total factor productivity estimation: A practical review. *Journal of Economic Surveys, 26: 98-128*, 2012.