UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA

STATISTICA ECONOMIA E FINANZA AZIENDALE

TESI DI LAUREA

Tecniche di Regressione Robuste:

il caso Cotonificio Albini SpA

RELATORE: PROF.SSA LAURA VENTURA

LAUREANDA: PATRIZIA DETIMO

ANNO ACCADEMICO 2008-2009

AI MIEI GENITORI:

Ciò che un genitore canta
vicino alla sua culla,
accompagnerà un bimbo
per tutta la sua vita.

Ringraziamenti

Finalmente è finita! Non ci posso credere … è finita!!! E solo tu lo sai, Papà, quanto ho desiderato questo momento, che pensavo non arrivasse mai. E sono qui a ringraziare tutte le persone che mi sono state vicine in questi anni, sostenendomi ed incoraggiandomi ad andare avanti.

Grazie Mamma, per aver sempre creduto in me anche nei momenti più duri, in cui credevo di non farcela, ti voglio bene.

Grazie Papà, per tute le volte che mi sei venuto in sogno a regalarmi un po' di grinta per il giorno dopo, grazie.

Grazie Giuseppe, Tesoro mio, per avermi incoraggiato a non mollare, per essermi stato accanto e per l'infinita pazienza che hai avuto … ora puoi cantare e suonare quando e quanto vuoi.

Grazie Michele ed Elena, per avermi sopportata … ogni estate con i libri in mano.

Grazie Manu e Giusy, per aver aspettato con pazienza questo giorno. Ce l'ho fatta!

Grazie a Teresa e Raffaele, per esservi sempre preoccupati di me e per avermi trattata sempre come una figlia, grazie.

Grazie Federica, per l'amicizia profonda e sincera che mi hai donato e per la tua presenza nei momenti di studio intenso.

Grazie Atika e Simone, per la disponibilità avuta nei miei confronti e per avermi regalato momenti di divertimento indimenticabili.

Ringrazio inoltre, tutte le amiche del Collegio Mazza, tutti gli amici di università e tutti coloro che in questi anni mi hanno fatto dono della loro amicizia, rendendo gradevole e felice la mia permanenza a Padova ed ogni ricordo ad essa legato.

Grazie a tutti gli Amici e Colleghi di Bergamo che mi sono stati vicini in questo periodo intenso … da oggi posso uscire senza limiti di tempo, evviva!!!

E' doveroso ringraziare anche il Cotonificio Albini SpA, in particolare il Dott. Leonardo Mangili e la Dott.ssa Elena Maffeis per avermi dato la possibilità di svolgere al meglio questo lavoro.

Ringrazio, infine, la mia relatrice professoressa Laura Ventura per la disponibilità e cortesia dimostratami in questo periodo di tesi e per aver messo a mio servizio la sua professionalità e le sue competenze.

Patrizia Detimo

INDICE

INTRODUZIONE

In un mercato fortemente competitivo, caratterizzato da una contrazione del ciclo di vita dei prodotti e da una riduzione dei tempi di risposta alla richiesta di mercato, le aziende *logistic-oriented,* sono sempre più stimolate ad adottare sistemi di previsione della domanda, che possano permettere loro di definire una strategia industriale adeguata.

Quando un'azienda non adotta un corretto sistema di previsione, o lo applica in maniera incompleta, la naturale conseguenza è l'adattamento del livello di scorte di

prodotti finiti alla variabilità della domanda. Un'operazione, quest'ultima, dannosa per l'azienda in quanto il costo di mantenimento del magazzino può salire fino al 30-40% del valore dei prodotti a stock, o addirittura rischiosa quando il capitale immobilizzato risulta particolarmente elevato e le scorte sono soggette ad obsolescenza, ovvero a svalutazione economica.

In questa tesi viene preso in considerazione il caso specifico di un'azienda tessile: il Cotonificio Albini S.P.A. che progetta e produce, a catalogo e su commessa, tessuti per camiceria.

L'attuale contesto competitivo internazionale presenta grandi sfide per il settore tessile: il fenomeno forse oggi più evidente è la fortissima concorrenza dei paesi asiatici, Cina in testa, che porta ad una pressione estrema sulla riduzione dei costi. Questi nuovi concorrenti non si limitano a perforare il mercato di bassa qualità, ma stanno sviluppando una presenza preoccupante anche laddove il Cotonificio Albini è leader incontrastato, vale a dire manifatture di alta qualità.

Queste sfide si vanno ad aggiungere alle criticità che da sempre caratterizzano il mercato dell'abbigliamento in generale, e della moda in particolare, contraddistinto da forte variabilità della domanda, stagioni brevi con conseguente rapida obsolescenza dei prodotti, grande imprevedibilità dei gusti e delle preferenze dei clienti. Per rispondere a queste sfide, le aziende dispongono di molteplici alternative possibili, alcune radicali, quali la delocalizzazione della produzione, che comporta elevati costi e rischi di insuccesso. Vi sono al contempo strumenti complementari, che hanno un effetto di miglioramento incrementale e di ottimizzazione dei processi esistenti, ma che mantengono i loro benefici anche in caso di decisioni radicali. Si tratta di interventi sulla previsione della domanda che a fronte di costi decisamente ridotti, permettono di conseguire molteplici vantaggi. Come prima cosa, una migliore previsione della domanda, soprattutto in un contesto turbolento come quello della moda, consente di migliorare la propria capacità di risposta al mercato, contenendo allo stesso tempo i costi, permettendo di pianificare meglio la produzione e riducendo le scorte di prodotti finiti. In secondo luogo, una migliore previsione consente di effettuare acquisti e gestire le scorte di materie prime in modo migliore, aumentando la disponibilità e riducendo gli obsoleti. Per migliorare l'economicità della gestione e

per la realizzazione di un management efficiente, le previsioni sono quindi un elemento primario.

Questa tesi discute alcuni aspetti di un'analisi di regressione per la previsione delle vendite al caso aziendale del Cotonificio Albini SPA. La tesi prevede lo studio critico del modello di regressione attualmente utilizzato dall'azienda per il calcolo della previsione e la proposta di un nuovo semplice modello, sempre nell'ambito della regressione lineare, che migliori quello attualmente utilizzato.

La struttura della tesi è la seguente. Dopo il primo Capitolo, in cui viene presentata l'azienda, il tipo di prodotto e la sua gestione, nel Capitolo 2 viene presentato il sistema attuale di previsione dell'azienda, con riferimento ad una stagione passata (Autunno/Inverno 2008-2009), in modo da poter confrontare i valori previsti con quelli reali, valutando così l'affidabilità del modello attualmente utilizzato. Sulla base dei risultati ottenuti e con il vincolo dell'esigenza dell'azienda di continuare ad utilizzare un modello regressione lineare (per la semplicità di analisi e comprensione), che si possa integrare facilmente con i loro programmi (Microsoft Excel e Microsoft Access), nel Capitolo 3 viene presentata la teoria della regressione robusta, come modello alternativo alla regressione lineare attualmente utilizzata dall'azienda. Infine, nel Capitolo 4 vengono analizzati e confrontati in modo puramente applicativo i dati stimati con il nuovo modello di regressione. I risultati che emergono, seppur preliminari, sono interessanti.

CAPITOLO 1

Il Cotonificio Albini SpA

"L'impegno e l'ambizione di produrre i tessuti più belli del mondo, perché ogni mattina la scelta della camicia è un piacere che esprime anche un po' di noi stessi".

(S.Albini)

Obiettivo di questo capitolo è presentare l'azienda, il settore e il mercato di riferimento. L'attenzione si focalizza soprattutto sul tipo di prodotto e sulla sua gestione, nonché sugli obiettivi di questa tesi e dell'azienda stessa.

1.1 Il settore e l'azienda

Il Cotonificio Albini è un'azienda specializzata nella produzione di tessuti per camiceria. Fondata nel 1876 ad Albino, in provincia di Bergamo, è sempre stata un'azienda a conduzione famigliare: si è, infatti, oggi alla quinta generazione.

La scelta della località di Albino, situata nella media Valle Seriana, è da ricondursi principalmente a quattro motivazioni strategiche, molto importanti a quell'epoca per lo sviluppo di un'azienda tessile:

- la presenza nell'area del più importante bacino di sviluppo del comparto cotoniero italiano e quindi la possibilità di interagire con aziende offerenti servizi e materie prime necessarie alla produzione tessile;
- l'esistenza di un elevato numero di società elettriche, dovuta all'abbondanza di risorse idriche, che potevano soddisfare il fabbisogno sempre crescente di energia. Le risorse idriche, inoltre, sono molto importanti anche per la produzione tessile, in relazione alle lavorazioni di tintoria e finissaggio;
- la possibilità di avvalersi della grande esperienza e competenza nel settore, maturate a livello di piccole manifatture da sempre presenti nella zona;
- la presenza di un'abbondante bacino di manodopera a cui attingere.

E' proprio in queste terre che, nella seconda metà dell'ottocento, si è sviluppata la prima industria cotoniera italiana ed ancor oggi, nelle valli intorno a Bergamo, vi è una delle più alte concentrazioni di industrie tessili d'Europa .

Il settore tessile cotoniero, in generale, può essere definito come un settore maturo, in quanto lo spazio per l'innovazione, soprattutto per la tipologia dei filati, è piuttosto limitato.

E' per questo che, essendo un ambiente molto competitivo, si presenta la necessità di agire su altre leve di differenziazione, quali il colore del tessuto, i disegni e le tecniche di finissaggio, che possono dare ai tessuti caratteristiche molto particolari.

Questo settore risente, inoltre, di andamenti ciclici legati alla moda ed alla situazione economica generale. La moda influenza oltre che la scelta dei colori e la tipologia dei disegni (righe piuttosto che quadri o tinte unite), anche la scelta dei consumatori finali nel preferire prodotti sostitutivi alla camicia. Si è verificato infatti, nel 1999, un calo delle vendite per il fatto che è stata preferita la T-shirt alla camicia, come sottogiacca o indumento per il tempo libero, causando una diminuzione delle richieste.

In periodi di crisi economica, poi, tutto il settore ne risente, poiché la camicia di alta qualità viene considerata un bene di lusso ed è quindi superfluo.

Nel 1992 l'azienda ha acquisito due prestigiosi marchi anglosassoni: Thomas Mason e David & John Anderson, destinati alla produzione di tessuti di fascia alta ed altissima basati su filati ritorti. Con il secondo marchio, in particolare, avviene la produzione del tessuto con le migliori qualità tecnicamente ottenibili da un tessuto di cotone, utilizzando titoli di filato estremamente sottili e raffinati.

Grazie a questa operazione commerciale, il Cotonificio Albini S.p.a. si garantisce una maggiore penetrazione sui mercati anglosassoni e di lingua inglese (Fig.1.1).

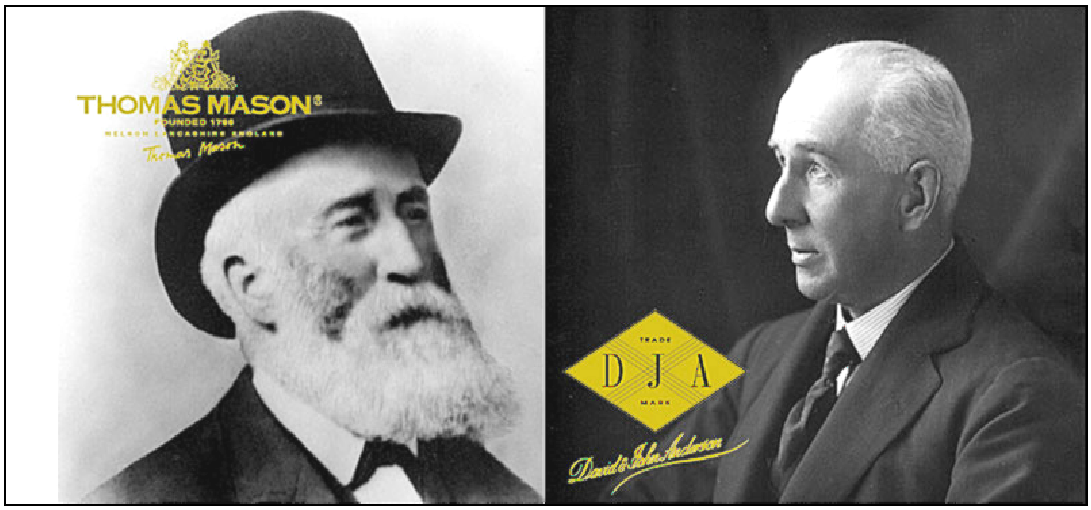


Fig.1.1: I marchi anglosassoni acquisiti dal Cotonificio Albini.

Oltre al cambiamento delle politiche commerciali, nella storia recente dell'azienda si evidenzia una grande propensione ad investire: nella innovazione dei prodotti, nelle tecniche produttive ed in quelle gestionali. Per far fronte alla crescente domanda da parte del mercato si provvede costantemente ad incrementare le capacità produttive, non solo per quanto riguarda l'attività di tessitura, ma anche interiorizzando attività, quali la tintoria dei filati e la nobilitazione.

Il Cotonificio Albini ha raggiunto gli obiettivi sopra ricordati attraverso la costituzione nel tempo di società controllate e/o unità produttive di divisione (rami d'azienda). Vediamo brevemente le principali tappe di questo processo.

Nel 1998 viene creata, non lontano dallo stabilimento principale, una piccola unità produttiva denominata "Albini 2". Lo stabilimento è interamente dedicato alla produzione di campionature. Tale attività si rivela essere sempre più cruciale perché incide sul momento in cui l'azienda propone alla clientela, in anteprima, la nuova collezione. La tempestività è fondamentale: da essa dipende l'afflusso degli ordini da parte dei clienti.

Nello stesso anno vi è stata la creazione di una divisione dedicata alle lavorazioni di finissaggio presso Brebbia (VA). Questo rappresenta un passo molto importante per l'interiorizzazione di una fase produttiva sempre più cruciale e per la creazione di valore aggiunto in termini di qualità e servizio al cliente. Nel corso degli ultimi anni,

l’impianto è stato completamente ammodernato e reso conforme alla tipologia produttiva che caratterizza il Cotonificio Albini, sia per tipologia di prodotto trattato che per qualità richiesta.

Nel 2000 il Cotonificio Albini ha acquisito la manifattura di Albiate S.p.a, società di antica costituzione (1836), localizzata in Brianza, ed operante nello stesso settore di riferimento.

Nel 2002, è avvenuta l’acquisizione della Dietfurt SRO, società con sede in Repubblica Ceca.

L’acquisizione di una partecipazione di controllo (50%) nel capitale della tintoria TFIL, avvenuta nel 2003, ha consentito al gruppo di far fronte al sempre crescente fabbisogno di filato colorato, dovuto all’ampliamento delle capacità produttive e ad un mercato che richiede una varietà di colori sempre più ampia.

Nel 2004, infine, ha iniziato ad operare l’ultima nata del Gruppo Albini: la divisione di Mottola, localizzata a Taranto.

1.2 Il prodotto e la sua gestione

Il Cotonificio Albini si è soprattutto focalizzato e distinto per la qualità e l’innovazione del prodotto.

Il prodotto finale è il tessuto in cotone utilizzato dai clienti per il confezionamento di camicie per uomo e per donna. Tale tessuto viene realizzato partendo dal filato greggio o, in minima parte, tingendo del tessuto greggio d’acquisto.

Tutti i tessuti vengono raggruppati ai fini commerciali in due macro linee:

- linea Albini;
- linea Thomas Mason.

Questi due raggruppamenti sono definiti “linee commerciali” e contengono delle varianti di prodotto destinate a diversi segmenti di mercato e quindi a clienti

differenti. La suddivisione in sottolinee commerciali facilita la gestione del cliente e la documentazione degli ordini. Infatti, ogni linea è affidata alla responsabilità di un agente che presenta la collezione ad un numero ristretto di clienti.

Ogni linea commerciale è formata da famiglie, ossia raggruppamenti di prodotti aventi alcune caratteristiche tecniche in comune, come ad esempio la tipologia di filato o il numero di battute a telaio. Ogni famiglia, infine, è composta da diversi articoli e ogni articolo da diverse varianti/colore (Tab.1.1).

LINEA	FAMIGLIA	ARTICOLO	VARIANTE
Linea Albini o Linea Thomas	70603	F33344	000002 000003 ...
		F33345	000056 000048 ...
	40503	F32356	000025 000028 ...
		F32357	000026 000015 ...
	...	...	...

Tab. 1.1: Esempio di gestione del prodotto: dalla linea alla variante.

Ogni sei mesi viene proposta una collezione con circa ottomila nuovi prodotti. Le due collezioni dell'anno fanno riferimento alle due stagioni: primavera-estate (P/E) ed autunno-inverno (A/I). Le presentazioni avvengono alla *Shirt Avenue* di Como e alla *Premier Vision* di Parigi, nella seconda metà di febbraio per la collezione primavera-estate dell'anno successivo e a fine settembre per la collezione autunno-inverno dell'anno successivo.

Ogni stagione viene identificata da un codice di tre cifre: le prime due indicano l'anno della collezione, mentre la terza indica la stagione (1 se è P/E, 2 se è A/I). Ad esempio, a metà febbraio 2007 è stata presentata la collezione P/E dell'anno 2008, ossia la collezione 081, mentre a fine settembre la collezione A/I 2008, ossia la collezione 082.

I clienti delle manifestazioni di Como e Parigi possono eseguire i primi esami, valutazioni e scelte dei tessuti su cartelle e *books*. Nelle cartelle vengono inseriti pezzi di tessuto che permettono di valutare le strutture dei tessuti (articolo), le disegnature (armature) e le variantature (colori). I *books*, come dice il nome, sono libri in cui sono contenuti vari campioni di tessuto, riuniti per tipologia di articolo e per linea di creazione.

Una volta esaminati gli articoli di proprio interesse, ciascun cliente sceglie i tessuti che gli serviranno per la preparazione dei suoi campionari, da presentare alle fiere dell'abbigliamento.

Dei tessuti scelti, i clienti richiederanno sia "tirelle" (fazzoletti di tessuto), per una maggiore percezione della qualità del tessuto, che "tagli", ossia campioni che corrispondono alla metratura necessaria per la creazione di una camicia di prova.

Dopo la fase di invio delle tirelle e dei campionari ai vari clienti si inizierà, sulla base degli ordini richiesti, la fase di produzione vera e propria della collezione.

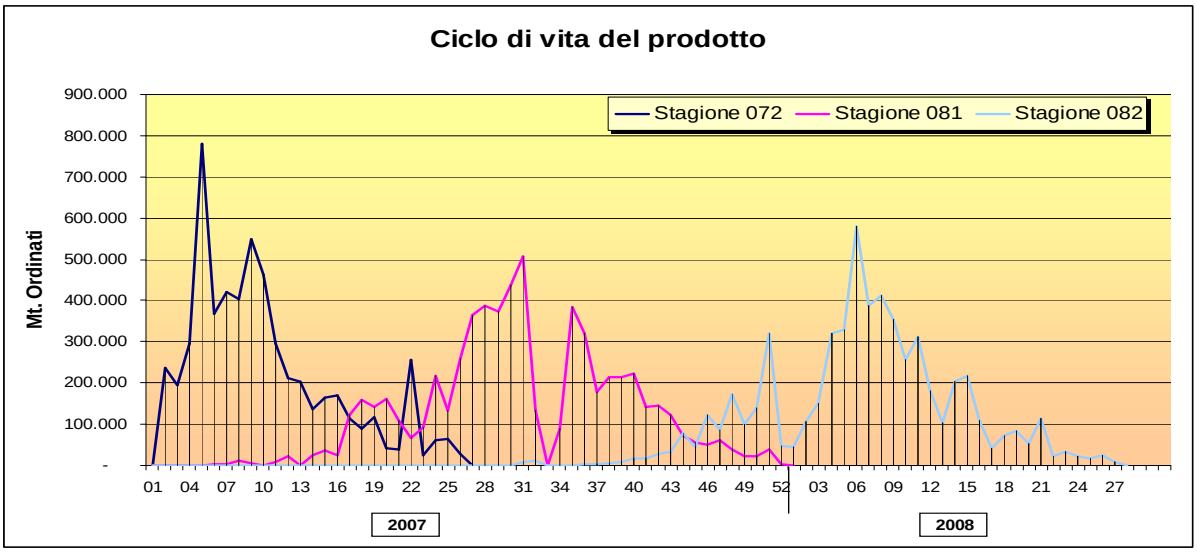


Fig. 1.2: Ciclo di vita del prodotto.

La Fig. 1.2 illustra mostra il ciclo di vita del prodotto relativamente agli ordini di produzione. Il grafico mostra l'andamento settimanale delle stagioni negli anni 2007 e 2008: la stagione 072 si riferisce alla collezione autunno/inverno 2007; la stagione 081 si riferisce alla collezione primavera/estate 2008; infine, la stagione 082 si riferisce alla collezione autunno/inverno 2008.

Per meglio capire le fasi del ciclo di vita del prodotto, nella Tab. 1.2 si mostra lo sfasamento temporale dallo studio del tessuto al capo finito negli anni 2006 e 2007.

La filiera del tessile: *dallo studio del tessuto al capo finito*	ANNO 2006												ANNO 2007											
	gen	feb	mar	apr	mag	giu	lug	ago	set	ott	nov	dic	gen	feb	mar	apr	mag	giu	lug	ago	set	ott	nov	dic
Elaborazione della collazione di tessuti	P/E '07 (**071**)		A/I '07-'08 (**072**)							P/E '08 (**081**)					A/I '08-'09 (**082**)							P/E '09 (**091**)		
Presentazione collezione		P/E '07							A/I '07-'08					P/E '08							A/I '08-'09			
Invio cartelle																								
Consegna Tagli e Tirelle			P/E '07							A/I '07-'08					P/E '08							A/I '08-'09		
Progettazione collezioni di abbigliamento				P/E '07												P/E '08								
Presentazione collezioni						P/E '07							A/I '07-'08					P/E '08						
Raccolta ordini dai distributori						P/E '07							A/I '07-'08					P/E '08						
Primi ordini di produzione tessuti						P/E '07							A/I '07-'08					P/E '08						
Produzione tessuti									P/E '07				A/I '07-'08						P/E '08					
Consegne ai confezionisti									P/E '07						A/I '07-'08						P/E '08			
Produzione capi										P/E '07							A/I '07-'08				P/E '08			
Consegne ai distributori													P/E '07						A/I '07-'08					

Tab.1.2: Sfasamento temporale dallo studio del tessuto al capo finito (anno 2006 e 2007).

L'innovazione e la varietà del prodotto, la tecnologia e la qualità, però, non bastano per affrontare efficacemente un mercato così complesso. I clienti hanno, infatti, la necessità di richiedere vari servizi: rispetto e rapidità dei tempi di consegna, capacità di gestire ordini frammentati, campionature e disegni personalizzati.

Per queste motivazioni quindi sono stati fatti degli sforzi per ottenere un'ampia scelta di tessuti pronti a magazzino da consegnare direttamente al cliente. Infatti, oltre ai prodotti nuovi gestiti su commessa e con un *lead time* di circa 3 mesi, sono messe a disposizione circa 2000 varianti di prodotto di servizio pronto (S*ervice Program*) già in parte presenti a magazzino o per i quali si assicura un tempo di consegna massimo di 60 giorni, poiché a magazzino è già pronto il filato colorato.

Per garantire invece la personalizzazione, sono stati creati tessuti esclusivi, ossia delle varianti di prodotto create su richieste specifiche del cliente, al quale si può garantire non solo l'unicità del disegno, ma anche del colore del filato utilizzato.

Per la gestione delle numerose varianti di prodotto si utilizza un codice definito "stato commerciale", che identifica se una variante appartiene alla tipologia produttiva di collezione, dei *Service Program*, delle esclusive o se è una variante di fuori collezione. Lo stato commerciale permette di classificare tutte le varianti attive (in giacenza o in produzione) in base all'anzianità della variante stessa e quindi al livello di obsolescenza. Inoltre, pilota le diverse modalità di programmazione e gestione della produzione: per esempio una variante di *Service Program* può avere un'urgenza maggiore rispetto a una variante di collezione. Nella Fig. 1.3 si nota la percentuale degli ordini di produzione divisa per stato commerciale, relativamente alla stagione autunno/inverno 2008 (082).

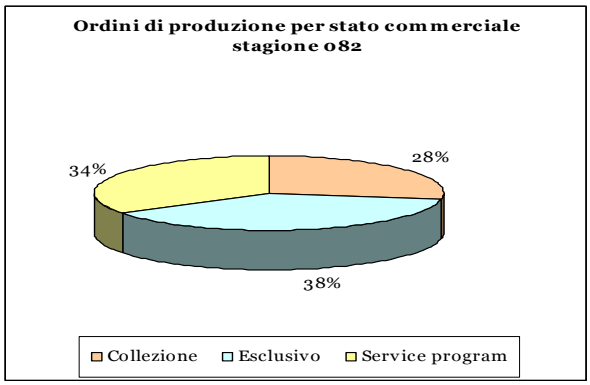


Fig. 1.3: Ordini di produzione per stato commerciale.

1.3 Il mercato di riferimento

La specializzazione dell'Azienda è nei tessuti tinti in filo, più complessi rispetto ai tessuti tinti in pezza, che rappresentano il 10% dell'intera produzione. Questi ultimi,

peraltro, sono presenti nel campionario per completezza di offerta e, di regola, sono acquistati esternamente allo stato grezzo, tinti e commercializzati.

Il prodotto deve avere caratteristiche di:

- Innovazione stilistica;
- Qualità;
- Prezzo;
- Servizio.

Per un produttore tessile in paesi ad alto costo del lavoro quali l'Italia, il prodotto in tutte le sue caratteristiche, come sopra indicato, rappresenta il più importante fattore di successo. Consapevoli di ciò, nell'azienda le risorse destinate allo sviluppo delle collezioni, alla realizzazione e alla preparazione di campioni per i clienti sono ingentissime. Più del 10% delle persone impiegate in azienda è costantemente dedicato a questa attività.

I tre segmenti di mercato che si ritengono più importanti sono:

- Tessuti per camicia classica, prevalentemente in filato ritorto, in cui si cerca di abbinare la raffinatezza dei disegni alle caratteristiche qualitative;
- Tessuti per camicia moda, caratterizzati da innovazioni raffinate nei disegni, colori, finissaggi per un modo di vestire internazionale;
- Tessuti per camicie sportive per un uso prettamente da tempo libero.

La materia prima utilizzata è in gran parte filato di cotone e lino (per la stagione estiva), solo in misura marginale si adoperano anche altre fibre (lycra, nylon, seta, acciaio).

Il mercato dei tessuti fini per camiceria è un mercato molto globalizzato, in cui relativamente pochi produttori di tessuti per camicia di tale fascia forniscono i clienti di tutto il mondo, e si posizionano nelle fasce di prodotto alte del mercato.

I costi di trasporto sono relativamente bassi rispetto al valore dei tessuti. Pertanto questi possono essere spediti in ogni parte del mondo. Il Cotonificio Albini esporta in

più di 65 paesi. Questo ha permesso di diversificare geograficamente i rischi ed allo stesso tempo di mantenere la specializzazione.

I principali paesi in cui si esporta sono: Inghterra, Spagna, Hong Kong, Germania, Francia, U.S.A., Giappone.

I clienti possono essere segmentati in due grandi categorie:

- i confezionisti di camicie;
- la distribuzione organizzata.

Fra i confezionisti di camicie, il Cotonificio Albini serve in ogni paese coloro che, per marchio, per qualità intrinseca del prodotto, per capacità di differenziarsi stilisticamente si posizionano nelle fasce alte e medio-alte di prezzo. Fra i clienti più importanti in questo segmento troviamo:

- In Italia: CIT-Giorgio Armani, Zegna, Dama S.p.a., Etro, Corneliani;
- In Germania: Hugo Boss, Seidensticker Logistick, Jacques Britt, Van Laack;
- In U.K.: Dewhirst Ltd., Thomas Pink Ltd., Paul Smith, Turnbull & Asser Ltd.;
- In Francia: Façonnable, Cinq Huitiemes Sa, Comptoir Int.De La;
- In U.S.A.: Nordstrom Inc., Polo Ralph Lauren, Ike Behar, Individualized Shirt, Gitman.

Per avere successo con questi clienti sono fondamentali:

- la differenziazione del prodotto: la presentazione, ossia, di prodotti continuamente nuovi coerentemente con l'evoluzione della moda di cui spesso i clienti sono protagonisti;
- il servizio al cliente, in termini di possibilità di studiare disegni esclusivi, di fornire i campioni nei tempi dovuti, di rispetto delle date di consegna della produzione, di produzione di lotti limitati e frazionati, di *quick response* ai mutamenti continui del mercato stesso;

- un rapporto qualità-prezzo accettabile, comunque non troppo alto, in quanto questi clienti devono a loro volta affrontare costi di marketing molto alti, quali pubblicità, royalties, etc.

L'altra importante categoria di clienti alla quale il Cotonificio Albini si rivolge, il cui peso è sempre più elevato, è quella rappresentata dalla distribuzione organizzata: ossia i clienti che sono in grado di controllare la propria distribuzione organizzando loro stessi la produzione, saltando così stadi di intermediazione e fornendo, pertanto, un migliore rapporto qualità-prezzo ai propri clienti finali.

Fra i clienti più importanti in questo segmento, troviamo:

- In Italia: Zanolini, Coin;
- In Spagna: Massimo Dutti, Cortefiel, Zara, El Corte Ingles;
- In Francia: Alain Figaret;
- In U.K.: Marks & Spencer, Turnbull & Asser, Hilditch & Key, Thomas Pink;
- In U.S.A.: Nordstrom, Brooks Brothers, Banana Republic.

Anche in questo caso i fattori di successo si rivelano essere: prodotto, servizio e rapporto qualità-prezzo. Questi clienti richiedono un livello di servizio ancora più alto e la possibilità, in particolare, di consegnare grandi lotti di qualità molto elevata e consistente in tempi brevi. Il potere contrattuale di questi clienti è molto elevato e crescente, in quanto è molto grande la loro capacità di mettere in concorrenza i vari produttori e di acquisire tessuti e prodotti finiti in ogni parte del mondo, sfruttando le condizioni più favorevoli.

In ogni caso, anche i confezionisti, con i propri marchi, gli stilisti e le case di moda sempre più stanno sviluppando una loro distribuzione diretta, fattore che sta diventando sempre più determinante per un maggior controllo del valore aggiunto di tutta la catena.

I concorrenti più forti, presenti in ogni parte del mondo, sono alcune imprese italiane ed europee, anche se, sempre di più, si affacciano nuovi protagonisti da paesi emergenti, quali la Turchia, l'India, il Pakistan e la Cina. Le imprese di questi paesi,

grazie ai costi molto più bassi, basano la loro competizione quasi unicamente sul prezzo e rappresentano una temibile minaccia alla competitività ed alla redditività delle aziende europee.

Nel mercato europeo il Cotonificio Albini S.p.A. si può stimare rappresenti circa il 10% della produzione totale nel suo segmento di prodotto.

I maggiori concorrenti sono i seguenti, in ordine di fatturato:

- Tessitura Monti, Italia;
- Getzner, Austria;
- Somelos, Portogallo;
- Cotonificio Honnegger, Italia;
- Testa S.p.A., Italia;
- Emmanuel Lang, Francia .

I produttori europei, e fra questi, in primis, gli italiani, sono gli unici al mondo che, sistematicamente, propongono ogni sei mesi collezioni nuove a cui dedicano ingenti risorse di sviluppo. Il principale obiettivo delle aziende europee è quindi presentare buone collezioni che si differenzino e siano migliori di quelle dei concorrenti. Avere un buon prodotto, in modo costante, ogni sei mesi, garantisce buona parte del successo.

1.4 Obiettivi dell'azienda

Obiettivo principale del Cotonificio Albini è essenzialmente rimanere leader del mercato nelle fasce di prodotti di media-alta qualità. Per poter perseguire tale obiettivo è necessario agire in due direzioni: prodotto e servizio.

La prima direzione consiste nel continuare la strada intrapresa di innovazione e qualità del prodotto, cercando però, in aggiunta, di far riconoscere ai clienti l'effettiva superiorità rispetto ad altri cotonifici concorrenti, in quanto i prezzi dei tessuti sono, in genere, meno competitivi.

Per quanto riguarda il servizio, invece, gli sforzi dovranno essere maggiori in quanto è necessario arrivare ad un ciclo produttivo più rapido e meglio organizzato per offrire maggiore puntualità e inferiore tempo di ciclo.

E' per questo che, da qualche anno, l'azienda ha implementato un sistema di previsione della domanda per meglio pianificare gli acquisti di materie prime, verificare l'effettiva capacità produttiva e garantire così ai propri clienti un livello di servizio migliore e tempi di consegna più brevi.

Nel prossimo Capitolo verrà presentato nel dettaglio il sistema di previsione adottato dall'azienda. L'attenzione è rivolta al tipo di modello utilizzato, alla valutazione della bontà del modello e all'analisi degli scostamenti tra le previsioni ottenute e i dati reali per una particolare stagione presa ad esempio.

CAPITOLO 2

Il sistema attuale di previsione dell'azienda

Per rispondere alla domanda: "A quanto ammonterà il venduto (quantità) dei nostri prodotti nella stagione?" da alcuni anni è stato implementato in azienda un sistema di previsione, basato su un modello di regressione lineare che, settimana dopo settimana, calcola, per ogni famiglia, la previsione a finire (ordini di produzione totali) in riferimento alla stagione in corso.

Obiettivo di questo Capitolo è presentare l'attuale sistema di previsione dell'azienda, su una stagione passata (A/I 2008-2009: codice 082) per i soli articoli di "collezione" ed "esclusivi". Lo scopo è presentare e studiare l'affidabilità del modello attualmente utilizzato.

Le analisi riportate di seguito sono state eseguite con il free software statistico ***R*** *(disponibile all'indirizzo www.r-project.org) che consente di eseguire analisi grafiche e calcoli statistici grazie ad alcune funzioni contenute all'interno di sue librerie. Tramite i pacchetti di base e quelli disponibili gratuitamente in rete,* ***R*** *consente ad esempio di stimare modelli lineari e modelli lineari generalizzati, di eseguire regressioni non lineari, di analizzare serie storiche e di effettuare test parametrici e non-parametrici.*

2.1 Il modello e i dati

Si ricorda che (si veda il Paragrafo 1.2) il tessuto ordinato dai clienti è suddiviso in tre (macro) categorie, individuate dalle tre variabili:

- TG: tagli;
- RT: tirelle;

- PDZ: veri ordini di tessuto, sul quale si vuole fare previsione.

Le prime due variabili fanno parte dei cosiddetti ordini di "campionario", grazie ai quali il cliente può farsi un'idea più precisa di come è fatto il tessuto, nonché creare una camicia di prova per la sua esposizione. La terza variabile, invece, rappresenta la fase di produzione vera e propria dell'azienda, dove il cliente ordina considerevoli metri di tessuto per creare la sua collezione.

Ogni singolo articolo viene, inoltre, classificato in base allo stato commerciale in: collezione, esclusivo o *service program*. Gli articoli di collezione e gli esclusivi vengono rinnovati ogni anno, mentre quelli di *service program* sono articoli standard riproposti ad ogni collezione.

Il modello di previsione di seguito illustrato riguarda solo articoli di "collezione" ed "esclusivi". Gli articoli di *service program* sono caratterizzati da una dipendenza temporale e viene adottato un metodo diverso di previsione, che non è oggetto di studio in questa tesi.

L'aggregazione dei dati è fatta a livello di famiglia (unità statistiche) e la numerosità campionaria n varia ad ogni stagione, essendo composta dal numero di famiglie presentate in fiera in una determinata stagione. Usualmente l'ordine di grandezza di n è compreso tra 300 e 350.

La previsione della variabile d'interesse PDZ viene effettuata secondo un aggiornamento settimana per settimana, utilizzando dei coefficienti stimati sulla base dei dati noti dalla stagione precedente, con un modello di regressione lineare della forma

$$Y = ß_{1t}X_{1t} + ß_{2t}X_{2t} + ß_{3t}X_{3t} + \varepsilon_t \qquad (2.1)$$

dove

- Y è un vettore di dimensione n relativo ai metri di produzione totali entrati a fine stagione, di seguito indicato anche con DTOT;
- X_{1t} è un vettore di dimensione n relativo ai metri di produzione già entrati fino alla settimana t (consuntivo progressivo), di seguito indicato con PO;

- X_{2t} è un vettore di dimensione n relativo ai metri di tagli già entrati fino alla settimana t, di seguito indicato con TG;
- X_{3t} è un vettore di dimensione n relativo ai metri di tirelle già entrati fino alla settimana t, di seguito indicato con RT;
- $\underline{ß}_t = (ß_{1t}, ß_{2t}, ß_{3t})$ sono parametri di regressione ignoti;
- ε_t è il vettore di dimensione n dei termini di errore, per i quali si assume che $\varepsilon_{ti} \sim N(0,\sigma^2)$, $i=1, \ldots, n$, indipendenti;
- t indica la settimana di riferimento, con $t= 1, \ldots, T$, e T indica l'ultima settimana della stagione.

Per stimare $\underline{ß}_t = (ß_{1t}, ß_{2t}, ß_{3t})$ alla settimana t, nel modello di regressione (2.1), vengono utilizzati i dati noti dalla stagione precedente, alla stessa settimana e il valore della variabile Y (o DTOT) a fine stagione.

In questo caso DTOT è un vettore di dimensione n relativo ai metri di produzione totali entrati nella stagione precedente ed è fisso per ogni settimana in cui si stima il modello. Invece, X_{1t}, X_{2t}, X_{3t} rappresentano, rispettivamente, vettori relativi ai metri di produzione, tagli e tirelle entrati fino alla stessa settimana della stagione corrente.

Una volta stimati, prima di calcolare la previsione, i coefficienti sono sottoposti ad un vincolo, per evitare di ottenere previsioni minori di quanto già effettivamente ordinato o addirittura negative. I vincoli posti sono:

- per $ß_{1t}$: $\hat{ß}_{1t} = \max(\hat{ß}_{1t}; 1)$
- per $ß_{2t}$: $\hat{ß}_{2t} = \max(\hat{ß}_{2t}; 0)$
- per $ß_{3t}$: $\hat{ß}_{3t} = \max(\hat{ß}_{3t}; 0)$.

Una volta calcolate le stime dei coefficienti, tali valori vengono inseriti nel seguente modello

$$Y_t = \hat{ß}_{1t} X_{1t} + \hat{ß}_{2t} X_{2t} + \hat{ß}_{3t} X_{3t}. \qquad (2.2)$$

Utilizzando quindi i valori di $X_{1t,} X_{2t}$ e X_{3t} nella settimana *t* della stagione corrente, si ottiene la previsione, Y_t, che dovrebbe essere sempre più precisa man mano che, settimana dopo settimana, ci si avvicina a fine stagione.

Un esempio della procedura appena presentata viene illustrato nel paragrafo seguente.

2.2 Sistema di previsione: Esempio sulla stagione 082, Autunno/Inverno 2008-2009

Lo studio di una stagione passata consente di poter confrontare i dati previsti dal modello adottato dall'azienda con quelli effettivamente registrati nella stagione e quindi di valutare la bontà del modello sulla base dei consuntivi. A tal fine sono stati estratti dal gestionale aziendale gli ordini progressivi settimanali per famiglia sulla stagione 082 (A/I 2008-2009) e sulla stagione precedente 072 (A/I 2007-2008).

Per meglio comprendere la struttura dei dati, nelle Tab 2.1 e 2.2 sono riportati, a scopo illustrativo, gli ordini progressivi per stagione, in riferimento alla settimana *t* = 40.

La Tab 2.1 si riferisce ai metri ordinati nella stagione 072 progressivi fino alla settimana 40: nelle colonne PO, TG ed RT è indicato l'ordinato progressivo fino alla settimana 40, mentre la colonna DTOT è relativa agli ordini totali entrati nella stagione per ogni famiglia.

Quindi, i metri ordinati di PO, TG ed RT variano ogni settimana, mentre DTOT rimane costante e rappresenta l'ordinato totale per famiglia sulla stagione a noi nota (072 – A/I 2007-2008). Tali dati ci permettono di stimare i coefficienti $\underline{ß}_t$ nella settimana in esame.

Famiglia	DTOT	PO	TG	RT
38793	1.335	1.200	350	1
38803	270	-	309	-
38813	-	-	9	-
38833	-	-	5	-
40503	24.345	-	313	15
40583	20.340	-	45	1
41033	5.530	-	120	-
41423	1.200	-	227	1
50083	27.270	-	46	-
50093	96.595	4.080	1.854	48
…	…	…	…	…

Stima dei coeff. da stagione 072

$$DTOT = ß_{1t}PO_t + ß_{2t}TG_t + ß_{3t}RT_t + \varepsilon_t$$

Tab. 2.1: Esempio input: ordini progressivi 072 (t= 40).

Una volta ottenute le stime di massima verosimiglianza di $\underline{ß}_t = (ß_{1t}, ß_{2t}, ß_{3t})$ si procede con il calcolo della previsione.

Famiglia	PO	TG	RT
38843	0	421,76	9,1
38853	0	49,5	0,78
38863	0	124	0,26
40503	0	334,9	2,21
40583	0	0	0
40873	0	0	0
40993	0	0	0
41033	0	60	0
…	…	…	…

Calcolo previsione su stag. 082

$$DTOT_t = \hat{ß}_{1t}PO_{1t} + \hat{ß}_{2t}TG_{2t} + \hat{ß}_{3t}RT_{3t}$$

Tab. 2.2: Esempio input: ordini progressivi 082 (t= 40).

La Tab. 2.2 si riferisce agli ordini progressivi alla stessa settimana 40 sulla stagione corrente (082). Si nota come l'input relativo alla stagione 082 non abbia la colonna relativa a DTOT, che va stimata sulla base del modello adattato ai dati della stagione 072.

In generale, sulla collezione autunno-inverno si iniziano a registrare i primi ordini a fine settembre, per poi chiudersi a fine giugno. Le settimane dell'anno interessate vanno dalla 39-esima fino alla 26-esima dell'anno successivo. Su questo range, è stata effettuata una regressione per ogni settimana e sono stati calcolati i coefficienti, sia puri che vincolati (Tab. 2.3).

		Coeff. Puri			Coeff. Vincolati		
week_ord	week	ß1_PO	ß2_TG	ß3_RT	ß1_PO	ß2_TG	ß3_RT
1	39	- 4,044	105,255	1.237,985	1,000	105,255	1.237,985
2	40	1,119	75,420	748,000	1,119	75,420	748,000
3	41	- 1,536	58,705	513,172	1,000	58,705	513,172
4	42	3,265	32,407	422,501	3,265	32,407	422,501
5	43	11,682	16,833	338,590	11,682	16,833	338,590
6	44	10,728	11,640	296,997	10,728	11,640	296,997
7	45	9,594	12,678	155,254	9,594	12,678	155,254
8	46	10,358	8,850	139,367	10,358	8,850	139,367
9	47	9,180	9,421	103,375	9,180	9,421	103,375
10	48	7,100	12,978	28,466	7,100	12,978	28,466
11	49	6,520	11,905	31,327	6,520	11,905	31,327
12	50	4,975	12,381	27,200	4,975	12,381	27,200
13	51	3,787	10,031	61,095	3,787	10,031	61,095
14	52	3,418	11,717	30,974	3,418	11,717	30,974
15	01	3,418	11,717	30,974	3,418	11,717	30,974
16	02	2,989	12,407	7,940	2,989	12,407	7,940
17	03	3,077	11,328	- 31,563	3,077	11,328	-
18	04	3,259	13,634	- 178,724	3,259	13,634	-
19	05	1,922	7,951	- 18,073	1,922	7,951	-
20	06	1,839	5,680	- 10,930	1,839	5,680	-
21	07	1,756	4,874	- 22,575	1,756	4,874	-
22	08	1,728	2,739	- 26,323	1,728	2,739	-
23	09	1,585	1,963	- 36,153	1,585	1,963	-
24	10	1,525	2,879	- 83,586	1,525	2,879	-
25	11	1,496	1,384	- 75,206	1,496	1,384	-
26	12	1,473	0,933	- 73,908	1,473	0,933	-
27	13	1,446	0,084	- 67,287	1,446	0,084	-
28	14	1,426	- 0,423	- 63,524	1,426	-	-
29	15	1,421	- 0,595	- 63,052	1,421	-	-
30	16	1,405	- 0,639	- 63,622	1,405	-	-
31	17	1,395	- 0,825	- 65,287	1,395	-	-
32	18	1,393	- 0,930	- 65,319	1,393	-	-
33	19	1,378	- 0,771	- 68,134	1,378	-	-
34	20	1,371	- 0,871	- 64,855	1,371	-	-
35	21	1,370	- 0,836	- 66,709	1,370	-	-
36	22	1,019	- 0,288	10,298	1,019	-	10,298
37	23	1,018	- 0,257	9,222	1,018	-	9,222
38	24	1,019	- 0,116	3,274	1,019	-	3,274
39	25	1,017	0,009	- 4,092	1,017	0,009	-
40	26	1,013	0,005	- 3,107	1,013	0,005	-

Tab. 2.3: Coefficienti di regressione stimati per ogni settimana sulla stagione precedente (p-value < 0.05).

Si nota come il passaggio dai coefficienti puri a quelli vincolati è importante soprattutto per la variabile RT, in quanto si ottengono delle stime negative dalla settimana "03" alla "26". Nelle Fig. 2.1-2.3 sono messi a confronto i coefficienti puri con quelli vincolati: per $ß_1$ la differenza è solo nelle prime settimane, mentre per $ß_2$ e $ß_3$ il passaggio ai coefficienti vincolati è fondamentale per correggere le stime.

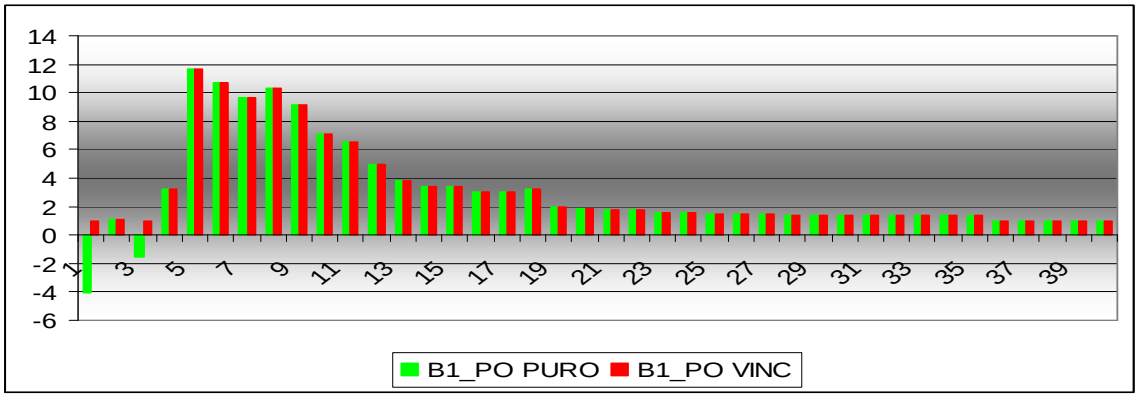


Fig. 2.1: Confronto tra coeff. Puro e coeff. Vincolato per la variabile PO.

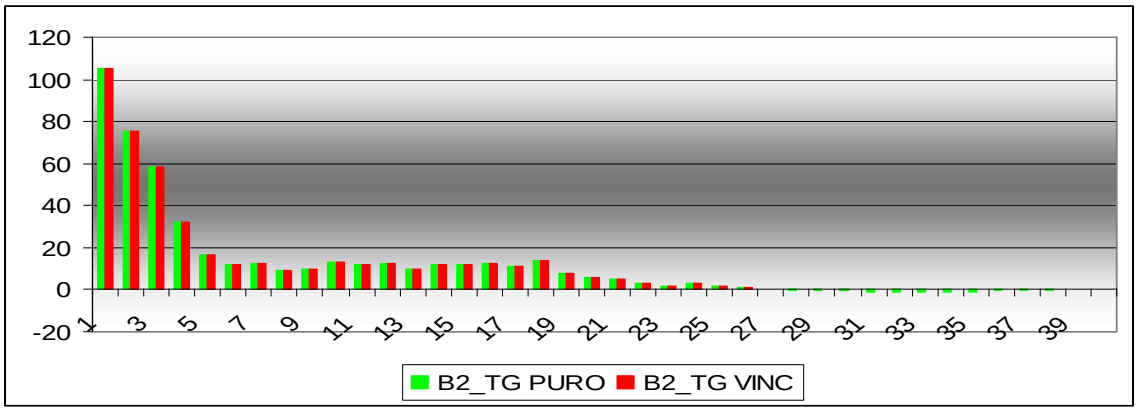


Fig. 2.2: Confronto tra coeff. Puro e coeff. Vincolato per la variabile TG.

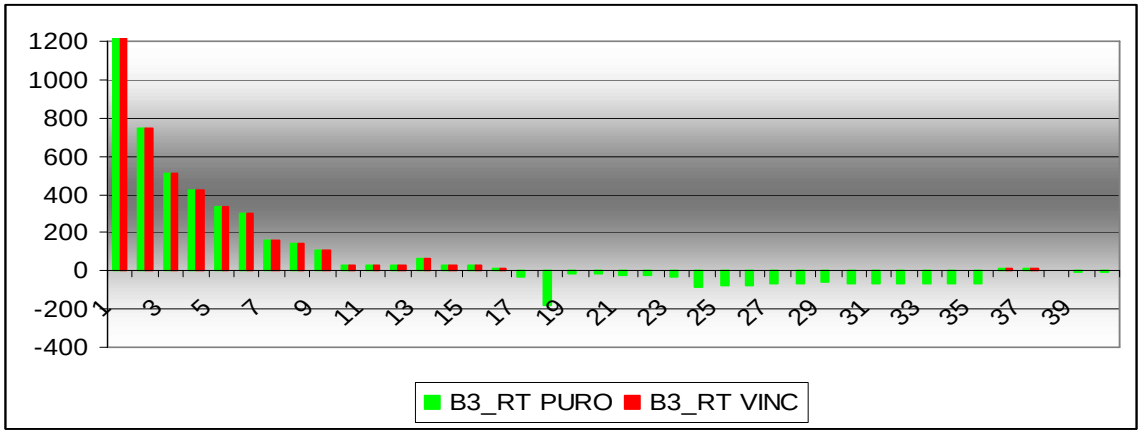


Fig. 2.3: Confronto tra coeff. Puro e coeff. Vincolato per la variabile RT.

Dopo la stima dei coefficienti, il passo successivo è calcolare la previsione della variabile *DTOT* per ogni settimana, utilizzando sia i coefficienti puri che quelli vincolati.

Nella Tab. 2.4 sono stati aggregati i dati per settimana (trascurando per ora la famiglia) e sono messe a confronto le previsioni con i consuntivi, calcolando il relativo errore di previsione. Sono state riportate in dettaglio le seguenti voci:

- il campo “week_ord” è stato inserito per comodità e ordina in senso crescente le settimane;
- i campi “DTOT_puro” e “DTOT_vinc” rappresentano le previsioni calcolate, rispettivamente, con coefficienti puri e vincolati, aggregate per settimana;
- il campo “DTOT_Act” rappresenta il consuntivo a fine stagione (ignoto in fase di previsione). Il dato è costante per tutte le settimane perché l’obiettivo è quello di “*prevedere quanti metri entreranno nella stagione in corso”*: quindi ogni previsione andrà confrontata con questo valore;
- i campi “Delta Puro vs Act” e Delta Vinc vs Act” rappresentano gli scostamenti calcolati come differenza tra la previsione e il consuntivo, sia per la previsione pura che per quella vincolata;
- i campi “Errore Puro vs Act%” e Delta Vinc vs Act%” rappresentano gli errori percentuali calcolati come rapporto tra delta e consuntivo.

StCom: Collezione & Esclusivo			Prev_pura	Prev_vincolata	Consuntivo a fine stg.	Scostamento vs consuntivo		Errore di previsione %	
Anno	week	week_ord	DTOT_puro	DTOT_vinc	DTOT_Act	Delta Puro vs Act	Delta Vinc vs Act	Errore Puro vs Act%	Errore Vinc vs Act%
2007	39	01	3.625.684	3.831.984	4.827.289	- 1.201.605	- 995.305	-24,9%	-20,6%
	40	02	3.776.487	3.776.487	4.827.289	- 1.050.802	- 1.050.802	-21,8%	-21,8%
	41	03	4.168.751	4.364.818	4.827.289	- 658.538	- 462.471	-13,6%	-9,6%
	42	04	3.961.374	3.961.374	4.827.289	- 865.915	- 865.915	-17,9%	-17,9%
	43	05	4.357.341	4.357.341	4.827.289	- 469.948	- 469.948	-9,7%	-9,7%
	44	06	4.741.511	4.741.511	4.827.289	- 85.778	- 85.778	-1,8%	-1,8%
	45	07	4.514.244	4.514.244	4.827.289	- 313.045	- 313.045	-6,5%	-6,5%
	46	08	5.350.016	5.350.016	4.827.289	522.727	522.727	10,8%	10,8%
	47	09	5.639.656	5.639.656	4.827.289	812.367	812.367	16,8%	16,8%
	48	10	6.065.442	6.065.442	4.827.289	1.238.153	1.238.153	25,6%	25,6%
	49	11	6.191.352	6.191.352	4.827.289	1.364.063	1.364.063	28,3%	28,3%
	50	12	5.797.561	5.797.561	4.827.289	970.273	970.273	20,1%	20,1%
	51	13	5.910.913	5.910.913	4.827.289	1.083.624	1.083.624	22,4%	22,4%
	52	14	5.751.436	5.751.436	4.827.289	924.147	924.147	19,1%	19,1%
2008	01	15	5.904.734	5.904.734	4.827.289	1.077.445	1.077.445	22,3%	22,3%
	02	16	6.019.483	6.019.483	4.827.289	1.192.194	1.192.194	24,7%	24,7%
	03	17	5.955.752	6.209.749	4.827.289	1.128.463	1.382.460	23,4%	28,6%
	04	18	6.397.622	7.836.891	4.827.289	1.570.333	3.009.602	32,5%	62,3%
	05	19	4.921.033	5.066.698	4.827.289	93.745	239.409	1,9%	5,0%
	06	20	5.400.117	5.488.267	4.827.289	572.828	660.978	11,9%	13,7%
	07	21	5.497.084	5.679.285	4.827.289	669.795	851.997	13,9%	17,6%
	08	22	5.611.818	5.824.503	4.827.289	784.529	997.214	16,3%	20,7%
	09	23	5.433.541	5.725.701	4.827.289	606.252	898.412	12,6%	18,6%
	10	24	5.335.371	6.011.033	4.827.289	508.082	1.183.744	10,5%	24,5%
	11	25	5.410.856	6.018.997	4.827.289	583.567	1.191.708	12,1%	24,7%
	12	26	5.427.480	6.025.683	4.827.289	600.191	1.198.394	12,4%	24,8%
	13	27	5.354.912	5.899.572	4.827.289	527.623	1.072.283	10,9%	22,2%
	14	28	5.452.831	6.043.136	4.827.289	625.542	1.215.847	13,0%	25,2%
	15	29	5.666.888	6.284.809	4.827.289	839.599	1.457.520	17,4%	30,2%
	16	30	5.686.510	6.317.431	4.827.289	859.221	1.490.142	17,8%	30,9%
	17	31	5.622.079	6.301.218	4.827.289	794.790	1.473.929	16,5%	30,5%
	18	32	5.659.876	6.358.987	4.827.289	832.587	1.531.698	17,2%	31,7%
	19	33	5.702.316	6.396.252	4.827.289	875.027	1.568.964	18,1%	32,5%
	20	34	5.727.688	6.413.609	4.827.289	900.399	1.586.320	18,7%	32,9%
	21	35	5.831.532	6.526.669	4.827.289	1.004.243	1.699.380	20,8%	35,2%
	22	36	4.901.351	4.954.210	4.827.289	74.062	126.921	1,5%	2,6%
	23	37	4.896.677	4.943.841	4.827.289	69.388	116.552	1,4%	2,4%
	24	38	4.897.491	4.918.740	4.827.289	70.203	91.451	1,5%	1,9%
	25	39	4.867.525	4.900.843	4.827.289	40.236	73.555	0,8%	1,5%
	26	40	4.866.095	4.891.393	4.827.289	38.806	64.104	0,8%	1,3%

Tab. 2.4: Analisi scostamenti di previsione vs consuntivo.

Come si nota dalla Tab. 2.4 l'errore è negativo nelle prime settimane e poi cresce rapidamente fino alla settimana 18 (colonna "week_ord"). L'errore poi si abbassa sensibilmente nella settimana 19 (periodo in cui l'azienda partecipa alla fiera Moda di Milano, inserendo parecchi ordini di Cartelle, Tagli e Tirelle), per poi risalire nelle successive settimane. Solo nelle ultime 5 settimane l'errore si stabilizza, quando tuttavia ormai gli ordini sono minimi e la stagione si può considerare chiusa.

Rappresentando graficamente l'errore percentuale (Fig. 2.4) si nota un andamento irregolare e l'utilizzo dei coefficienti vincolati non produce effetti migliorativi. Lo sbalzo nelle settimane 18 e 19 coincide con la partecipazione alla fiera "Milano Unica", in cui si registrano in modo massivo ordini di Cartelle, Tagli e Tirelle.

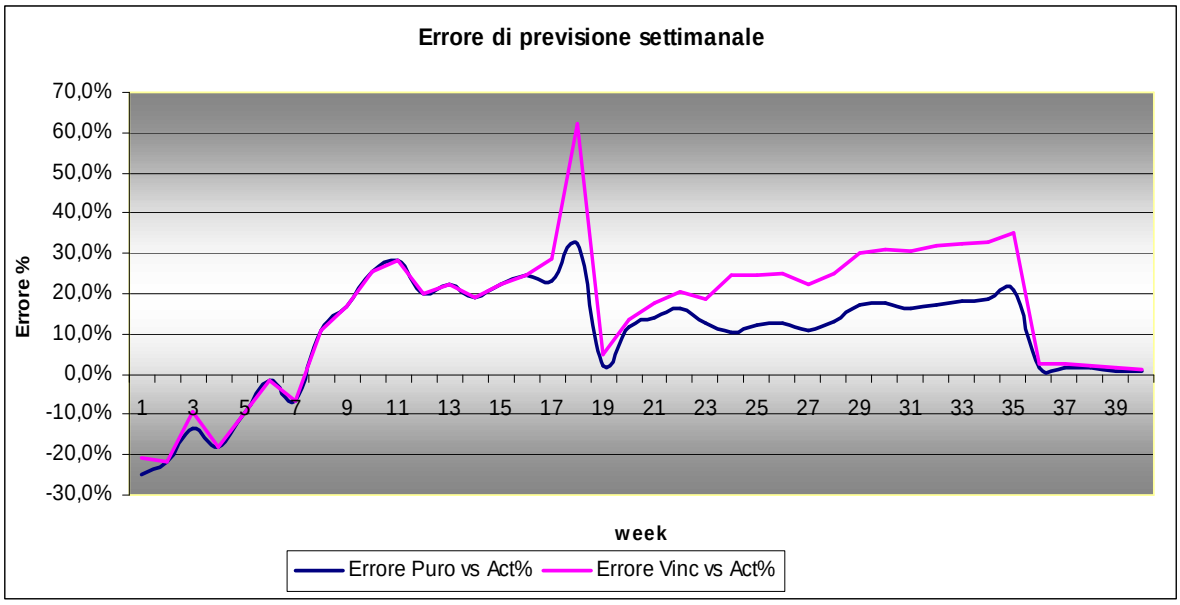


Fig. 2.4: Errore di previsione settimanale sia per famiglie di collezione che esclusivi.

Separando i due stati commerciali si ottengono le Fig. 2.5 e 2.6

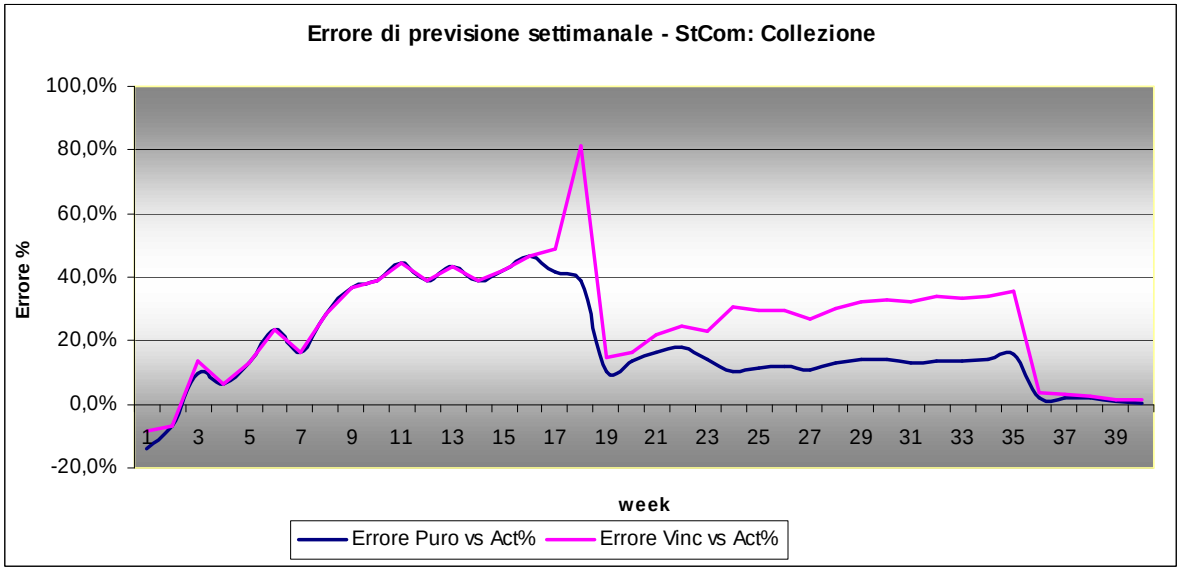


Fig. 2.5: Errore di previsione settimanale per famiglie di collezione.

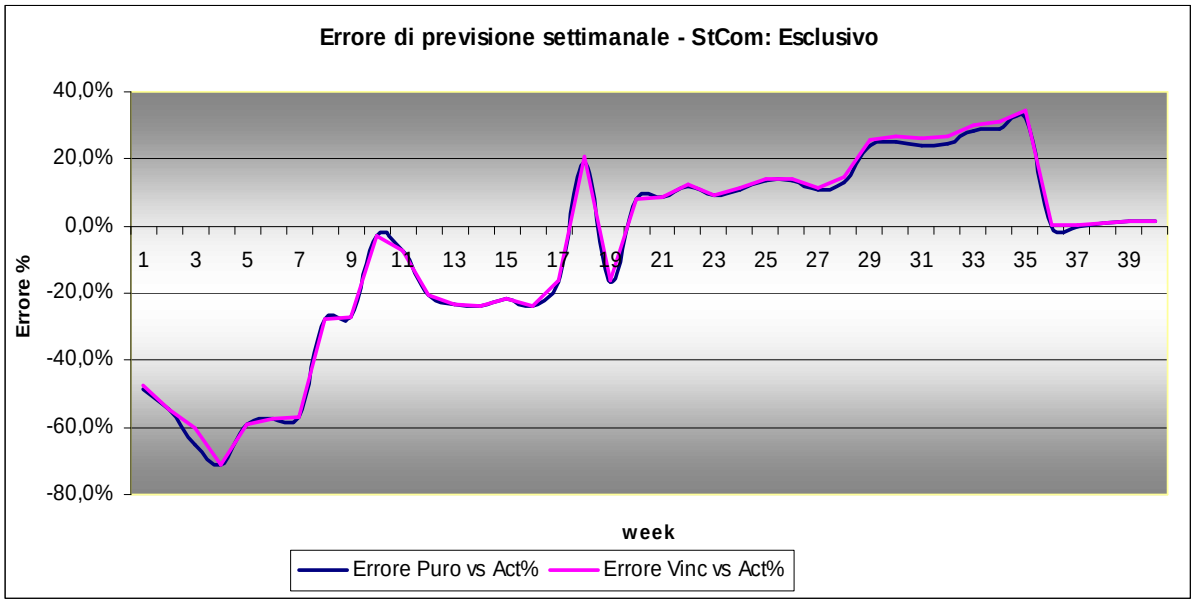


Fig. 2.6: Errore di previsione settimanale per famiglie esclusive.

Pur distinguendo tra collezione ed esclusivo, l'errore di previsione resta irregolare.

Mentre per la collezione si nota un andamento simile a quello complessivo (si veda Fig. 2.4), per lo stato commerciale esclusivo l'andamento è del tutto irregolare; infatti, gli ordini di questo tipo sono difficili da prevedere perché richiesti su specifica e in esclusiva dal cliente.

A livello di famiglia, la situazione non cambia. Le Fig. 2.7-2.10 mostrano l'andamento della previsione nelle varie settimane per le quattro famiglie più vendute.

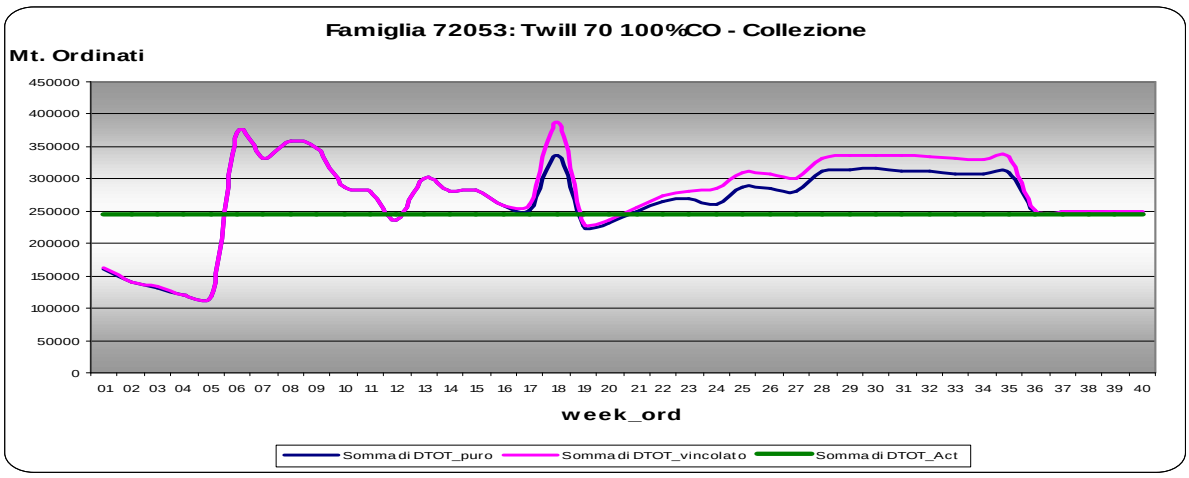


Fig. 2.7: Previsione settimanale dei metri ordinati per famiglia 72053.

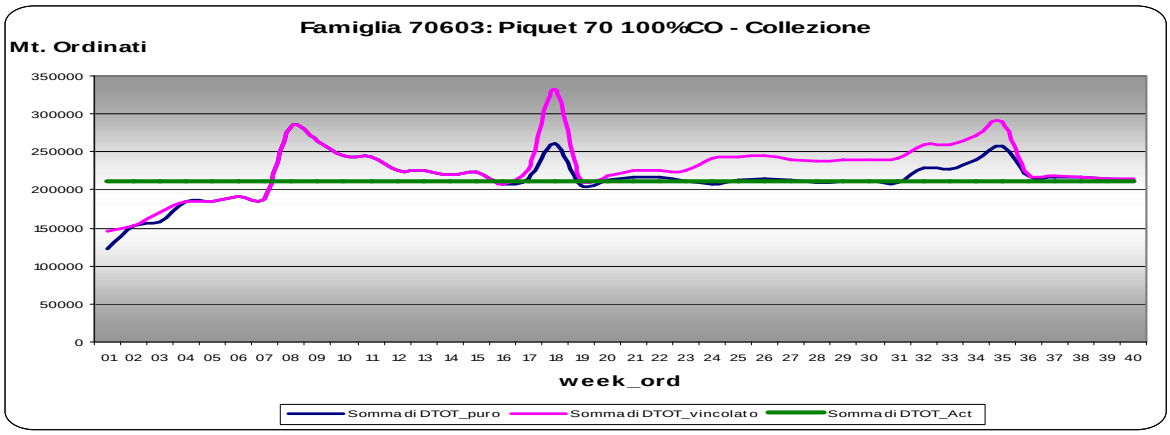


Fig. 2.8: Previsione settimanale dei metri ordinati per famiglia 70603.

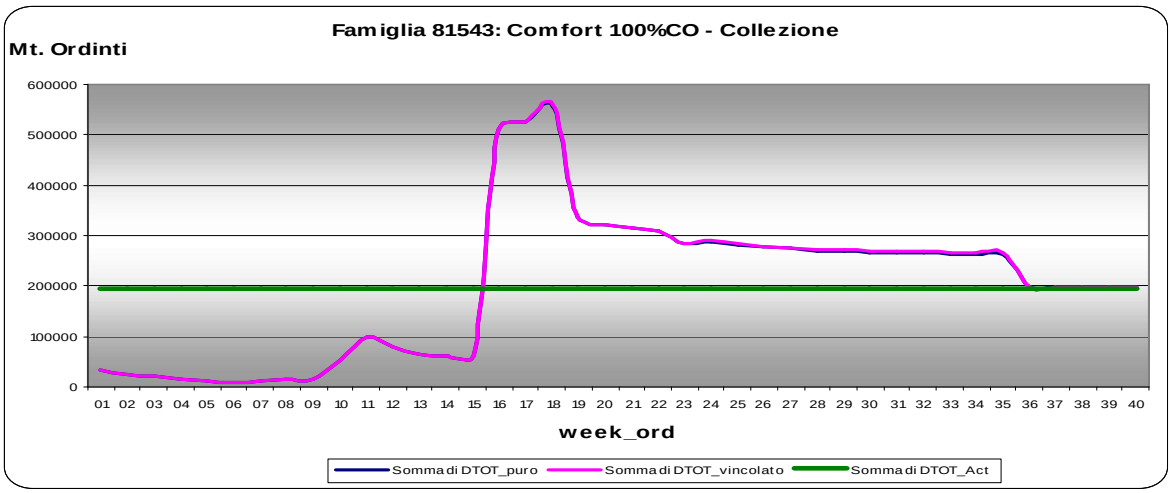


Fig. 2.9: Previsione settimanale dei metri ordinati per famiglia 81543.

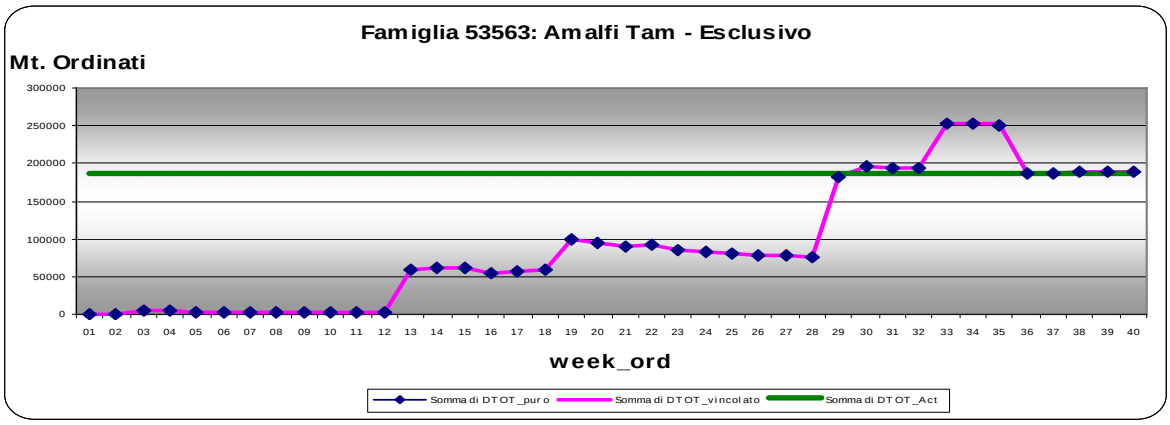
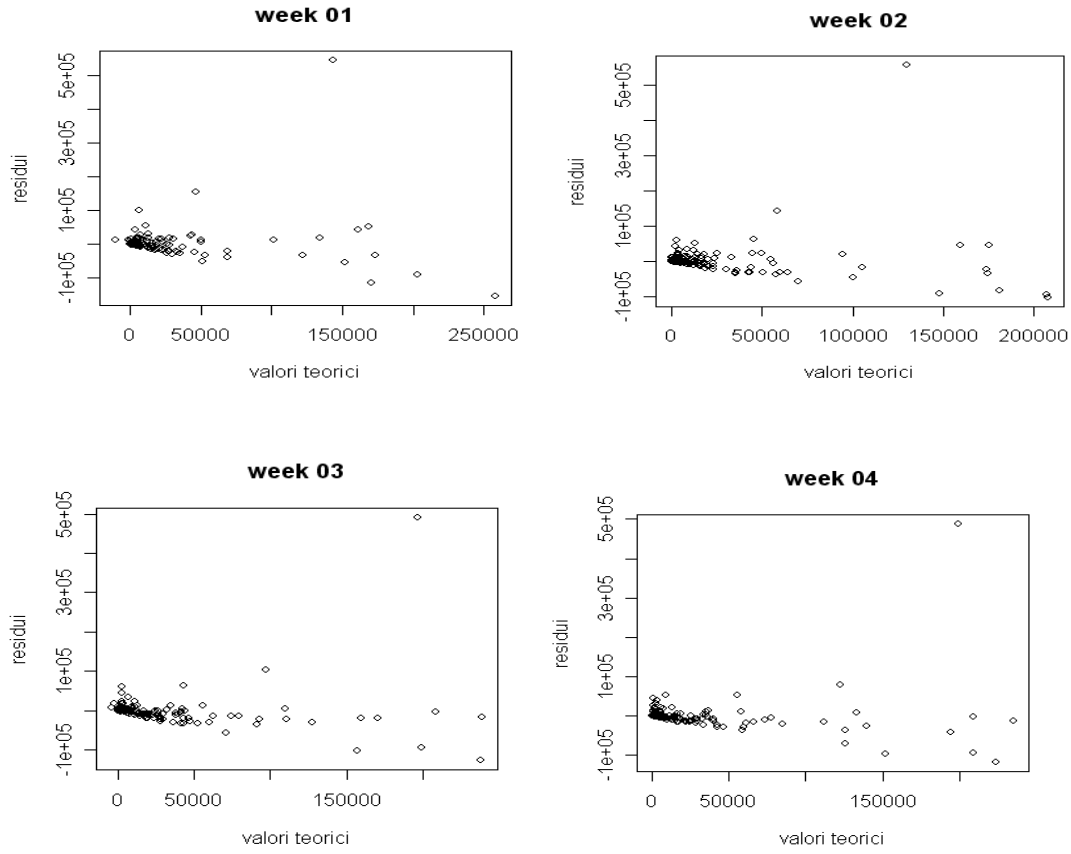


Fig. 2.10: Previsione settimanale dei metri ordinati per famiglia 53563.

2.3 Valutazione della bontà del modello

Al fine di valutare la bontà del modello adottato, sono state effettuate due analisi molto semplici: l'analisi dei residui e il calcolo del coefficiente di determinazione R^2. Lo strumento principale per il controllo empirico del modello è l'analisi dei residui, utile per evidenziare eventuali importanti scostamenti dagli assunti del modello (*Pace, Salvan*, 2001). Esiste un'ampia letteratura che tratta queste tecniche (cfr., Draper e Smith, 1981) e nel seguito, ci si limiterà a considerare semplici strumenti grafici. Per evidenziare eventuali non linearità un diagramma utile è il grafico dei residui rispetto ai valori stimati dal modello (si veda Fig. 2.11). Eventuali strutture o trend presenti nel grafico indicano che i residui non sono casuali e quindi che i regressori non hanno colto tutta la variabilità della variabile risposta.



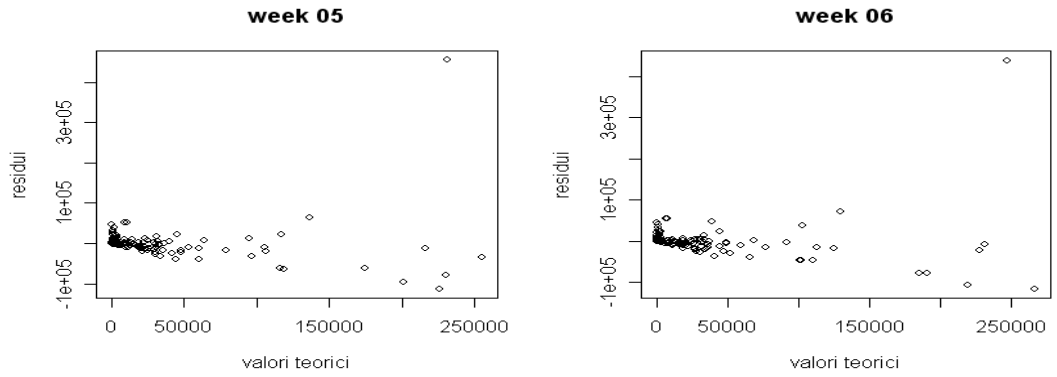
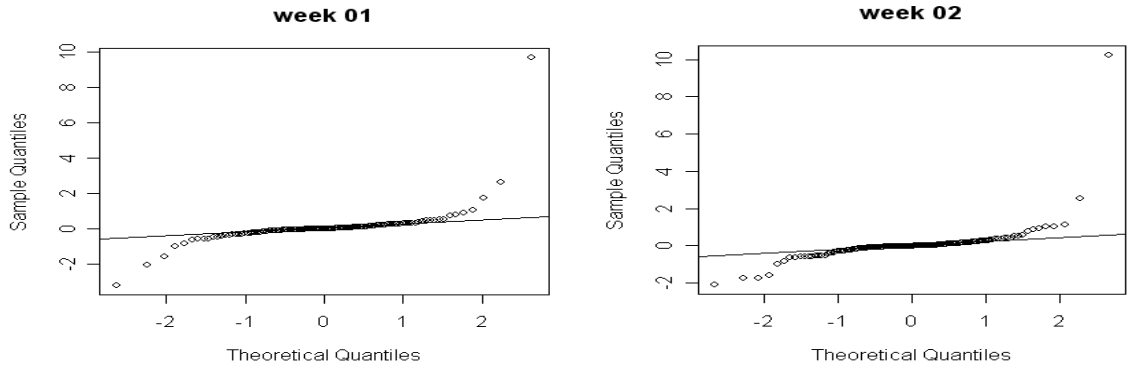


Fig. 2.11: Analisi dei residui rispetto ai valori stimati dal modello per alcune settimane.

Dalla Fig. 2.11 è possibile individuare la presenza di valori anomali (*outliers*) nei dati. I valori anomali sono osservazioni che si discostano dalla maggior parte dei dati osservati nella variabile risposta e che presentano residui elevati (*Bortot, Ventura, Salvan*, 2000).

Un altro test grafico utilizzato per la valutazione dell'ipotesi di normalità si può ottenere tramite il grafico delle probabilità normali per i residui, come mostrato in Fig. 2.12. Questo test pone a confronto i quantili empirici con quelli di una distribuzione normale. Se i quantili empirici sono in accordo con quelli di una distribuzione gaussiana, le coppie di punti si distribuirebbero lungo la retta di riferimento (*Iacus, Masarotto*, 2007).



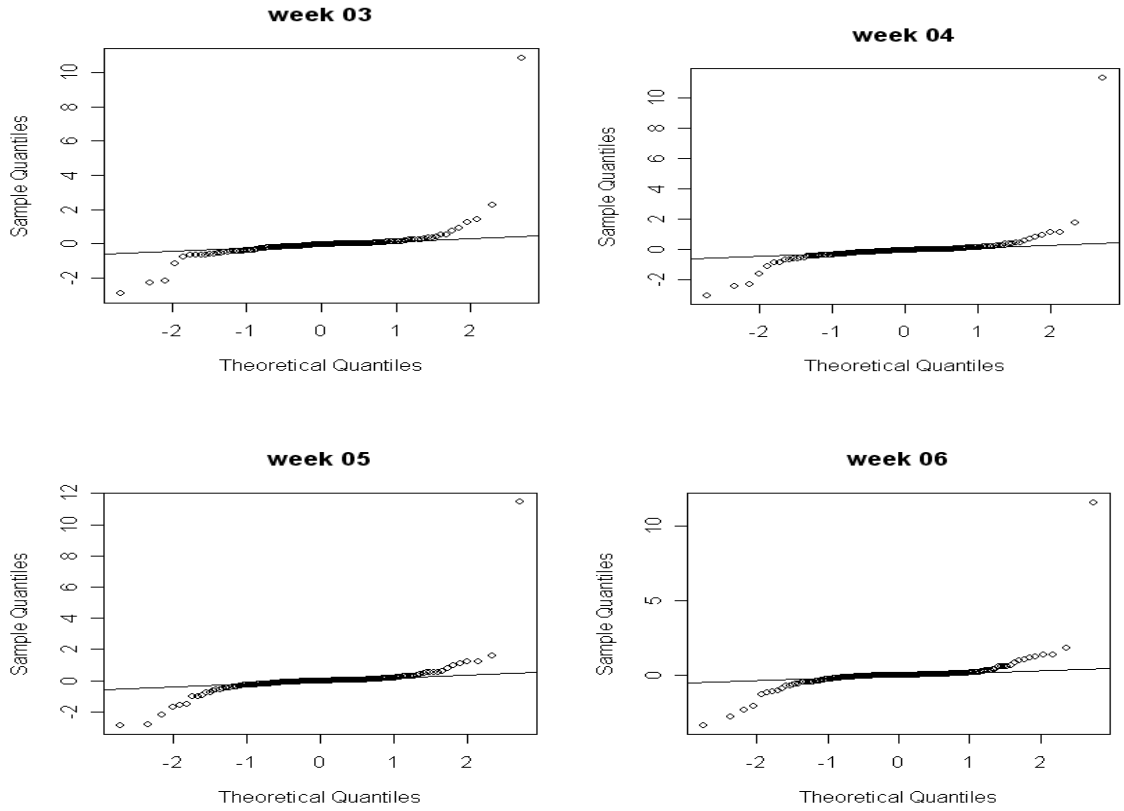


Figura 2.12: Verifica della normalità dei residui.

Lo scostamento dalla retta di riferimento fornisce la conferma che la distribuzione normale risulta inappropriata e l'allontanamento dalla normalità si nota soprattutto sulle code.

Infine, un indice utile ad esprimere la bontà del modello stimato è il coefficiente di determinazione R^2, dato da

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Il numeratore esprime la somma dei quadrati della regressione (SQR). Il denominatore, invece, è detto somma totale dei quadrati (SQT). Se il modello è un buon modello, l'SQR dovrebbe essere il più elevato possibile; quindi tanto più il coefficiente è vicino a uno (valore massimo), tanto più ci riterremo soddisfatti del modello. A tale scopo è stato calcolato per ogni settimana il suddetto indice rappresentato nella Fig. 2.13.

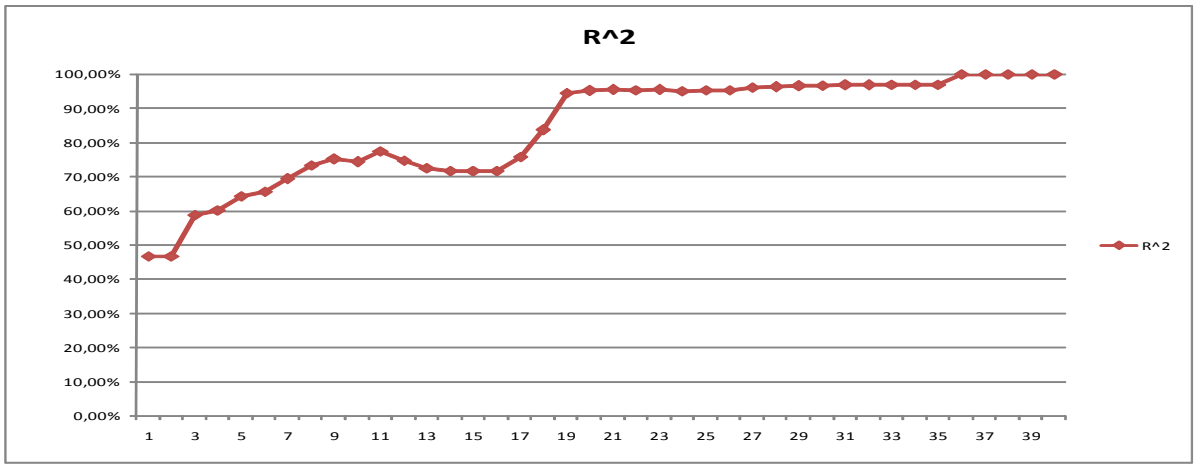


Fig. 2.13: Andamento dell'Indice di determinazione R^2.

Anche da questo indice si nota come la settimana 19 sia determinante. Infatti, a partire da questa settimana l'R^2 migliora.

2.4 Presentazione dei comandi in R

Il primo passo da compiere è caricare i dati in R sia in per la stagione passata (072) che per quella corrente (082) e per ogni settimana. Per la stagione 072 si hanno 40 file composti da 347 righe e 4 colonne: le righe si riferiscono alla famiglia di tessuto, mentre le colonne al tipo di tessuto ordinato (DTOT, PO, TG, RT). Per la stagione 082, invece, si hanno 40 file composti da 347 righe, una per ogni famiglia e 3 colonne (PO, TG, RT): manca infatti DTOT perché è la variabile ignota da stimare.

Per ogni settimana, il primo passo da compiere è caricare i dati relativi alla stagione passata:

```
> prec <-
read.csv2("C:/inputdati/072/01_072.csv",header=TRUE,
sep=";", quote="\"", dec=",", row.names=1, fill=TRUE)
> attach(prec)
```

In questo modo è stato caricato il file 01_072.csv, dove 01 indica la settimana che si sta caricando e 072 la stagione passata.

Il secondo passo è stimare i coefficienti del modello lineare (2.1) con il comando *"lm"* senza intercetta:

```
> fitprec<- lm(DTOT~ PO+TG+RT-1)
```

Questo comando permette di ottenere le stime dei coefficienti di regressione delle variabili esplicative che servono per la stima del modello sui dati della stagione corrente. Il passo successivo è quello di caricare i dati della stagione che si vuole prevedere:

```
> act <- read.csv2("C:/input dati/082/01_082.csv",
header=TRUE, sep=";", quote="\"", dec=",", fill=TRUE)
> attach(act)
```

Con il comando *predict* si calcola la previsione, utilizzando i coefficienti di regressione ottenuti in `fitprec`:

```
> dtot_01 <- data.frame(predict(fitprec,
newdata=data.frame(PO=act[,2], TG=act[,3], RT=act[,4]),
se.fit=T, , interval= "confidence"), row.names=act[,1])
```

Così si ottiene la previsione calcolata nella settimana 01 per ogni famiglia sui metri di tessuto che entreranno a fine stagione. Per comodità, i dati ottenuti sono stati esportati in Excel con il comando *write.table* :

```
> write.table(dtot_01,file="C:/input
dati/ACT/01_act.xls", append=FALSE, quote=TRUE, sep="\t",
dec=",", row.names=TRUE)
```

Per la valutazione della bontà del modello (si veda Paragrafo 2.3), l'analisi dei residui è stata effettuata con il seguente comando:

```
> plot(fittedprec, rstandard(fitprec), main="week 01")
```

Infine, il comando *qqnorm* applicato ai residui del modello riporta il diagramma quantile-quantile.

```
> qqnorm(fitprec)
```

La procedura appena presentata per il calcolo della previsione alla settimana 01 è uguale per tutte le altre settimane.

2.5 Conclusioni

In questo capitolo è stato esposto il metodo di previsione utilizzato attualmente dall'azienda per stimare la quantità di tessuto (metri) che verrà ordinata in una stagione. Tale metodo utilizza un modello di regressione lineare, le cui stime di massima verosimiglianza vengono opportunamente vincolate per evitare di ottenere previsioni negative o inferiori rispetto a quanto già ordinato.

Dal confronto tra le previsioni e i dati noti (si veda Tab. 2.4) è emerso che il modello produce previsioni sovrastimate. Neanche il passaggio alle stime vincolate produce effetti migliorativi.

La causa dell'errata specificazione del modello è legata (come si è visto nel Paragrafo 2.4) alla violazione degli assunti di base del modello di regressione lineare: la non normalità dei residui, la presenza di valori anomali e in alcuni casi l'indipendenza dei residui (Greco e Ventura, 2006)

Come vedremo nei prossimi capitoli, una possibilità in caso di scorretta specificazione del modello, oppure in presenza di elevati valori anomali (*autliers*), è ricorrere alla **statistica robusta** (Hampel et al., 1986). Essa costituisce un approccio alla statistica che ha lo scopo di salvaguardare rispetto a eventuali deviazioni delle ipotesi statistiche assunte.

CAPITOLO 3

La Regressione Robusta

Questo capitolo presenta una breve rassegna su alcune tecniche di regressione robuste.

Dopo aver introdotto il concetto di robustezza, l'attenzione si focalizza sugli stimatori di tipo M per parametri di regressione, che saranno oggetto di applicazione in questa tesi. Ulteriori approfondimenti sulla regressione robusta sono reperibili, per esempio, in Huber (1964, 1973) e Hampel et al. (1996), Maronna e Yohai (2006).

3.1 Introduzione alla teoria della robustezza

Le procedure di inferenza statistica sono basate sui dati solo parzialmente in quanto hanno anche grande importanza tutte le ipotesi che, più o meno esplicitamente, vengono formulate sul fenomeno sottostante. Tali ipotesi rappresentano il tentativo di formalizzare convenientemente da un punto di vista matematico conoscenze approssimative e, a volte, vere e proprie congetture sul fenomeno stesso. Anche nei casi più semplici, vengono spesso formulate ipotesi sull'indipendenza ed identica distribuzione delle osservazioni, sul meccanismo generatore dei dati (normalità, linearità, ...) o, in ambito bayesiano, sulle distribuzioni a priori sugli ignoti parametri.

La statistica parametrica classica, in particolare, a seconda delle informazioni disponibili, si basa sull'assunzione di un modello parametrico, specificato da

$$F = \{ f(y;\theta), \vartheta \in \Theta \}, \qquad (3.1)$$

dove $f(y;\theta)$ denota una funzione di densità e $\Theta \subseteq R^k$, $k \geq 1$, lo spazio parametrico. Il modello (3.1) copre un ruolo centrale nelle procedure classiche di inferenza.

Ora, se è vero che in molte branche della scienza i modelli matematici permettono di introdurre semplificazioni essenziali, è altrettanto vero che il loro uso dovrebbe essere giustificato in base ad una sorta di *principio di stabilità*: piccoli spostamenti dal modello ipotizzato non dovrebbero produrre grossi cambiamenti nelle conclusioni inferenziali finali. Infatti, il modello (3.1) può non rispecchiare esattamente la realtà, o per la presenza di dati anomali (*outliers*) nel campione osservato o per il carattere approssimato del modello teorico stesso.

Molte procedure classiche di inferenza sono pesantemente vincolate alla validità delle ipotesi formulate, non rispondendo così a questa sorta di principio di stabilità. Ciò è particolarmente vero in ambito parametrico, in quanto diverse procedure classiche risultano molto, ed a volte estremamente, sensibili a piccoli spostamenti della distribuzione dei dati dal modello assunto. Ad esempio, alcuni stimatori tradizionali classici, come la media aritmetica, sono generalmente molto sensibili alle osservazioni anomale, che quindi possono considerevolmente alterare il valore delle stime.

• *ESEMPIO 1: Stima della media (Hampel et al., 1986)*

Si considerino i dati di Cushny e Peebles (1905) sul prolungamento del sonno a seguito della somministrazione di due farmaci. Questi dati sono stati utilizzati da numerosi autori come esempio di campione casuale semplice da una distribuzione normale di media μ e varianza σ^2. Le n = 10 differenze degli effetti dei due farmaci per ogni individuo sono

y = (0.0, 0.8, 1.0, 1.2, 1.3, 1.3, 1.4, 1.8, 2.4, 4.6) .

È noto che la stima di massima verosimiglianza di μ è data dalla media campionaria

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = 1.58 .$$

Tuttavia, si nota che l'osservazione $y_{10} = 4.6$ si discosta dalla maggior parte dei dati e ciò risulta poco compatibile con l'ipotesi di normalità. Un modo naturale per studiare la sensibilità e stabilità di uno stimatore è quello di far variare una singola osservazione del campione y e quindi di esplorare l'effetto di tale variazione sulla stima. In particolare, per valutare l'effetto di una singola osservazione sul valore della stima di μ, si è fatta variare l'osservazione y_{10} nell'intervallo (0.0, 5.0) e, per ogni valore di y_{10} in questo intervallo, si è ricalcolato il valore di $\hat{\mu}$.

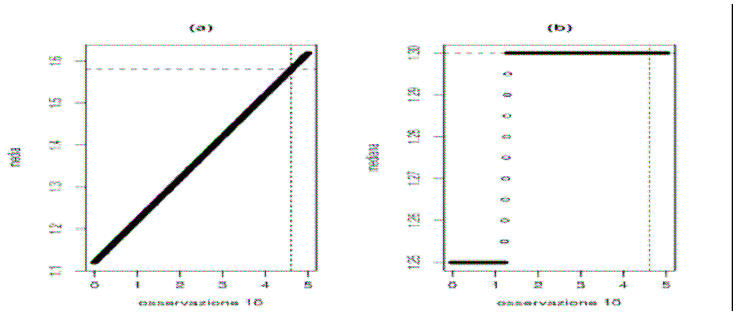


Fig. 3.1: Grafici dei valori della media campionaria (a) e della mediana (b) al variare dell'osservzione y_{10}.

I risultati di questa analisi sono riassunti nella Fig. 3.1(a) e mostrano come la media campionaria dipenda direttamente dal valore di una singola osservazione. Nella Fig. 3.1(b) è invece riportato il valore della mediana, stimatore alternativo per il parametro di posizione μ, al variare dell'osservazione $y_{10.}$ In questo caso si nota come il valore di questa stima di μ rimane limitato al variare di y_{10}. ◊

Quando una procedura inferenziale risulta molto sensibile rispetto a piccole deviazioni, sono allora necessari metodi statistici affidabili in un qualche intorno del modello. Una procedura statistica (nella pratica uno stimatore, una statistica test, etc.) poco sensibile a piccoli o moderati scostamenti (in termini della distribuzione dei dati) dal modello ipotizzato è detta ROBUSTA. La teoria della robustezza descrive le

proprietà delle procedure statistiche in un intorno del modello parametrico. Essa costituisce un approccio alla statistica, non un ramo della statistica, in quanto ha lo scopo di salvaguardare rispetto a eventuali deviazioni dalle ipotesi statistiche assunte.

3.2 La Robustezza

Si possono distinguere due concetti di robustezza:

1) robustezza rispetto alla contaminazione (*outliers*);

2) robustezza rispetto alla specificazione scorretta (*misspecification*).

Il primo concetto desidera tener conto, nel modello statistico e nell'inferenza, della possibile presenza nel campione di dati anomali, ossia di una qualche frazione di osservazioni che non sono effettivamente rappresentative della popolazione oggetto di studio, che invece è in accordo con il modello (3.1). Questi dati anomali possono essere causati da errori di rilevazione o di codificazione, ma anche determinati da una lieve eterogeneità della popolazione, riconducibile a distribuzioni con code pesanti. Anche i problemi conseguenti alla discretizzazione, provocata dall'arrotondamento dei valori o dalla riduzione in classi, possono essere trattati in questo ambito. In genere, i dati sono ritenuti di buona qualità se la frazione di dati anomali non supera l'1% (si tenga tuttavia in mente l'Esempio 1).

Il secondo concetto di robustezza desidera preservare le procedure di inferenza, basate su una specificazione convenzionale del modello statistico, dalla plausibile inadeguatezza del modello (3.1) come modello esatto. In altri termini, il modello statistico (3.1), pur catturando qualitativamente e quantitativamente aspetti importanti dei dati, in particolare quelli su cui si desidera fare inferenza, per il suo carattere approssimato non descrive con esattezza tutti gli aspetti della variabilità della popolazione.

Questi due diversi aspetti della robustezza, pur concettualmente distinti, sono tuttavia molto vicini e in molti casi equivalenti da un punto di vista pratico. Infatti, una

scorretta specificazione del modello può essere la causa della presenza nel campione osservato di dati anomali, ossia di osservazioni distanti dalla maggioranza dei dati.

Identificato un modello (3.1) che si adatti alla maggioranza dei dati osservati, obiettivo della statistica robusta è quello di individuare delle procedure atte a prevenire effetti, a volte disastrosi, causati da valori anomali molto distanti o da osservazioni influenti, ossia osservazioni che hanno un peso rilevante nella determinazione di una quantità di interesse. In particolare, la statistica robusta si occupa di determinare delle procedure di inferenza con buone proprietà (consistenza ed efficienza), necessarie per prevenire perdite di efficienza se i dati provengono proprio dal modello ipotizzato, e robuste rispetto a piccole e moderate deviazioni dal modello. Una formalizzazione dell'ampliamento del modello parametrico (3.1), necessario per pervenire a procedure inferenziali robuste sia rispetto alla contaminazione che alla specificazione scorretta, è basata sull'assunzione che il modello F sia una buona approssimazione - nel senso che sia in un qualche intorno della distribuzione - del vero modello che ha generato i dati, da cui può differire per alcuni aspetti. Obiettivo dell'inferenza resta comunque il parametro θ. La Fig. 3.2 riassume vari metodi per analizzare i dati e i diversi approcci per l'inferenza sul parametro θ. Sono raffigurati tre metodi per analizzare i dati: in (a) le procedure inferenziali classiche, in (b) i metodi basati sull'eliminazione delle osservazioni anomale e in (c) i metodi robusti con buone proprietà in termini di efficienza.

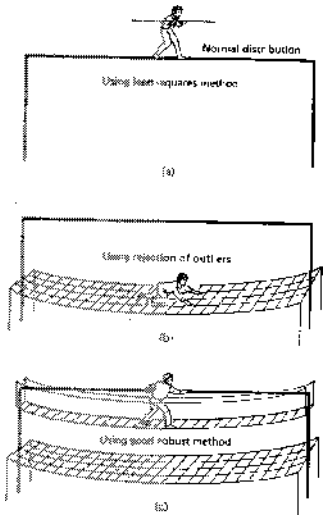


Fig. 3.2: Vari metodi per analizzare i dati (Hampel et al., 1986).

Le procedure parametriche classiche implicano l'efficienza, mentre quelle basate sull'eliminazione delle osservazioni anomale la stabilità. La statistica robusta si colloca pertanto in una situazione intermedia.

La principale critica mossa alla robustezza è che qualunque statistico accorto si muoverebbe adottando una qualche procedura di individuazione e eliminazione delle osservazioni anomale, procedendo poi, solo in un secondo momento, con l'applicazione delle procedure classiche adeguate per l'inferenza. Nella pratica questo modo di procedere può essere poco pertinente per vari motivi:

1. In molte applicazioni quando la numerosità del campione è elevata o le variabili in analisi sono numerose (come ad esempio in un modello di regressione con diverse covariate), l'individuazione delle osservazioni anomale è molto difficile, e nel caso delle osservazioni influenti, che sono quelle osservazioni con grande effetto nella stima del modello, è addirittura impossibile, a meno che non si faccia di fatto già ricorso ad una qualche tecnica robusta.
2. Il campione con le osservazioni anomale eliminate o con i dati arrotondati difficilmente è coerente con il modello ipotizzato. Ad esempio, verosimilmente, l'operazione di pulizia dei dati anomali comporta errori sia di eliminazione sia di conservazione (si eliminano dati che dovrebbero essere tenuti e/o si tengono dati che dovrebbero essere eliminati).
3. Infine, ma non meno importante, l'esperienza empirica ha mostrato che anche le migliori tecniche di eliminazione delle osservazioni anomale funzionano sempre meno bene di tecniche di inferenza robusta.

3.3 Stimatori di tipo M

Una scorretta specificazione del modello statistico, come ad esempio uno scostamento più o meno marcato dall'ipotesi di normalità, può portare, a seconda dei

casi, a pesanti conseguenze sulle procedure inferenziali. Pertanto, nelle situazioni in cui non si abbiano sufficienti informazioni sul fenomeno d'interesse, provenienti dai dati o da altre fonti (ad esempio indagini precedenti), è auspicabile ricorrere a statistiche robuste rispetto alla specificazione scorretta (*misspecification*). La letteratura statistica propone l'utilizzo di equazioni di stima robuste, ovvero opportune equazioni di stima che devono sottostare a determinati vincoli. Una funzione di stima è una generica funzione dipendente dal campione osservato $y = (y_1, \ldots, y_n)$ e dal parametro oggetto d'inferenza θ. Un'importante classe di stimatori legata a funzioni di stima robuste è quella degli stimatori di tipo *M* (*M-estimators*), introdotta da Huber (Huber, 1964, Hampel *et al*., 1986), che sarà oggetto di questa tesi. Il nome deriva da "*maximum likelihood type estimators*" essendo, gli stimatori di tipo *M*, una generalizzazione degli stimatori di massima verosimiglianza.

Si definisce uno stimatore di tipo *M* per θ uno stimatore $\tilde{\vartheta}$ che soddisfa l'equazione

$$\psi(Y;\vartheta) = \sum_{i=1}^{n} \psi(y_i;\vartheta) = 0\ , \qquad (3.2)$$

ove $\psi(.)$ è una funzione nota con valori in R^k. Nell'ambito del modello *F*, un'equazione di stima è detta non distorta se

$$E[\psi(Y;\vartheta)] = 0, \quad \forall \vartheta \in \Theta.$$

Tale condizione non implica la non distorsione dello stimatore $\tilde{\vartheta}$ ottenuto come soluzione della (3.2) (Desmond, 1997), bensì fornisce l'argomento principale per mostrare la consistenza di $\tilde{\vartheta}$.

Per semplicità di esposizione si assuma ϑ scalare e sia $l_*(\vartheta) = \sum_{i=1}^{n} \frac{\partial \log f(y_i;\vartheta)}{\partial \vartheta}$ la funzione *score* di verosimiglianza ottenuta dal generico elemento di F. Si dimostra (Godambe, 1960) che, sotto F , la funzione *score* gode di una proprietà di ottimalità fra tutte le funzioni di stima non distorte. Precisamente, se il parametro è scalare, vale

$$\frac{V(\ell_*(\vartheta))}{\{E[\partial \ell_*(\vartheta)/\partial\vartheta]\}^2} \le \frac{V(\psi(Y;\vartheta))}{\{E[\partial \psi(Y;\vartheta)/\partial\vartheta]\}^2}, \forall \vartheta \in \Theta,$$

Sia $Deff = \dfrac{V(\psi(Y;\vartheta))}{\{E[\partial\psi(Y;\vartheta)/\partial\vartheta]\}^2}$. Il numeratore del *Deff* è la varianza dell'equazione di stima, mentre il denominatore è un indice di sensibilità della funzione nel discriminare i valori di ϑ: tanto più concentrata è la funzione attorno a un valore ϑ^*, allora tanto più potente sarà nel discriminare valori in un intorno di ϑ^*, e ciò implica un *Deff* piccolo (Desmond, 1997).

Due proprietà interessanti dello stimatore $\tilde{\vartheta}$ ottenuto da un'equazione di stima non distorta della forma (3.2) riguardano il suo comportamento asintotico, in analogia alla teoria degli stimatori di massima verosimiglianza. Infatti, sotto tenui condizioni di regolarità, si può dimostrare che lo stimatore $\tilde{\vartheta}$ è consistente. Inoltre, sotto le stesse condizioni, vale l'approssimazione

$$\tilde{\vartheta} \approx N_k(\vartheta, V_\psi(\vartheta)),$$

ove $V_\psi(\vartheta)$ è una matrice definita positiva, data da

$$V_\psi(\vartheta) = M_\psi^{-1}(\vartheta) A_\psi(\vartheta) M_\psi^{-T}(\vartheta), \tag{3.3}$$

con $M_\psi(\vartheta) = E\left[\dfrac{\partial}{\partial\vartheta}\psi(Y;\vartheta)\right]$ e $A_\psi(\vartheta) = E\left[\psi(Y;\vartheta)\psi(Y;\vartheta)^T\right]$.

Per $k = 1$, una naturale stima della (3.3) è

$$\hat{V}(\hat{\vartheta}) = \frac{\sum_{i=1}^n \psi(\hat{\vartheta}; y_i)^2}{\left[\sum_{i=1}^n \dfrac{\partial}{\partial\vartheta}\psi(\vartheta; y_i)\Big|_{\vartheta=\hat{\vartheta}}\right]^2}. \tag{3.4}$$

Uno strumento che gioca un ruolo fondamentale nella teoria della robustezza è la funzione d'influenza. La funzione d'influenza per $\tilde{\vartheta}$ rispetto al modello F, calcolata nel punto c, indicata con IF(c; $\tilde{\vartheta}$), è definita come (Huber, 1981)

$$IF(c;\tilde{\vartheta}) = -M_\psi^{-1}(\vartheta)\psi(c;\vartheta), \qquad c \in \Re, \forall\,\vartheta \in \Theta. \tag{3.5}$$

La richiesta di robustezza si traduce nell'imposizione di opportune condizioni di limitatezza sulla IF, la più importante delle quali è rappresentata dalla richiesta che la quantità

$$\gamma = \sup_c \left\| IF(c; \tilde{\vartheta}) \right\|,$$

in cui $\|.\|$ denota la norma Euclidea, sia finita. L'indice γ è noto nella letteratura robusta come l'indice di sensibilità ai grandi errori (*gross error sensivity*). Se γ risulta limitato, allora si ha la B-robustezza di $\tilde{\vartheta}$ e per uno stimatore di tipo M , l'indice γ è finito se e solo se la funzione è limitata.

Naturalmente un'equazione di stima robusta apporta un notevole beneficio in termini di affidabilità delle procedure inferenziali quando il modello è effettivamente contaminato, ad esempio è formato dalla mistura di più modelli. Ciò non vuol dire, però, che i benefici di un'equazione di stima robusta siano incondizionati: la ricerca di robustezza si sconta in termini di efficienza dello stimatore ottenuto. Questo è il compromesso che Huber (Huber, 1981) sintetizza con la dicitura "*two-person zero-sum game*". Quindi un obiettivo da perseguire, nel momento in cui si decide di utilizzare un'equazione di stima robusta, può essere quello di ricercarne una che fornisca lo stimatore *B*-robusto ottimo, ossia lo stimatore con varianza asintotica minima nella classe degli stimatori con indice γ limitato. Definito un insieme $\varepsilon_a = \{\psi(.) | \gamma \le a, a \in \Re\}$, l'equazione di stima ottima in ε_a è tale per cui

$$tr(V_*(\vartheta)) < tr(V(\vartheta)), \forall \psi \in \varepsilon_a ,$$

con $\psi(.)$ equazione di stima *B*-robusta ottima tra le *B*-robuste.

• *ESEMPIO: Modello di Posizione e Stimatore di Huber*

Per un modello di posizione, ossia tale che $f(y;\theta) = f_o(y - \theta)$, è naturale scegliere la funzione $\psi(y;\vartheta)$ della forma $\psi(y;\vartheta) = \psi(y - \vartheta)$, con $\psi(.)$ funzione di stima nota. In questa situazione, l'espressione della IF e della varianza asintotica di uno stimatore di tipo M per un parametro di posizione si riducono a (*Hampel et al., 1986*)

$$IF(c;\tilde{\vartheta}) = \frac{\psi(c;\vartheta)}{\int \psi'(y-\vartheta) f_o(y-\vartheta) d\vartheta}$$

e

$$V(\vartheta) = \frac{\int \psi^2(y-\vartheta) f_o(y-\vartheta) d\vartheta}{\left[\int \psi'(y-\vartheta) f_o(y-\vartheta) d\vartheta\right]^2}.$$

Si noti anche che, se F_θ è assolutamente continua, tali espressioni per la IF, e la varianza asintotica mantengono la loro validità anche se la funzione $\psi(.)$ è derivabile a meno di un numero finito di punti.

Si consideri il caso particolare in cui (y_1, …, y_n) sono osservazioni indipendenti tratte da una normale *N(θ, 1).* La stima di massima verosimiglianza (SMV) $\bar{y}$ è soluzione di una equazione del tipo (3.2) con $\psi(y;\vartheta) = y - \vartheta$. La corrispondente IF non è limitata, essendo proporzionale a *y* - *θ.* Uno stimatore B-robusto, ossia robusto rispetto a contaminazioni infinitesimali, può essere definito sostituendo nell'equazione di stima (3.2) la funzione $\psi(y;\vartheta) = \psi_k(y-\vartheta)$ con

$$\psi_k(t) = t \min(1, k/|t|) = \begin{cases} -k & se \quad t < k \\ t & se \quad -k \le t \le k \\ k & se \quad t > k \end{cases} \qquad (3.7)$$

ove *k* è una costante positiva assegnata. Tale stimatore fu introdotto da Huber (1964) e gode di numerose proprietà interessanti. Ad esempio, esso ha varianza asintotica minima nella classe degli stimatori Fisher consistenti con γ^* limitato (*Hampel et al., 1986*). Una scelta comune di *k* è il valore 1.345 che permette di raggiungere il livello di efficienza desiderato in termini di robustezza. ◊

3.4 Regressione robusta

Si consideri il seguente modello di regressione

$$y_i = x_i^T \beta + \sigma\varepsilon_i, \; i = 1, \ldots, n, \tag{3.8}$$

dove y_i denota la variabile risposta, x_i è un vettore p-dimensionale di regressori, $\beta \in \mathbb{R}^p$ ($p > 1$) è l'ignoto parametro di regressione, $\sigma > 0$ è un parametro di scala e εi è un termine di errore con distribuzione nota $f_0(.)$ simmetrica attorno allo 0. I diversi tipi di dati anomali che si possono incontrare in una analisi di regressione lineare sono presentati nella Fig. 3.3.

Gli usuali stimatori ai minimi quadrati per il parametro β sono molto sensibili alla presenza di valori anomali o di osservazioni influenti nei dati osservati. Essi sono la soluzione dell'equazione

$$\sum_{i=1}^{n} r_i x_i = 0, \tag{3.9}$$

con $r_i = (y_i - x_i^T \beta)/\sigma$, che è del tipo (3.2) con funzione $\psi(.)$ non limitata.

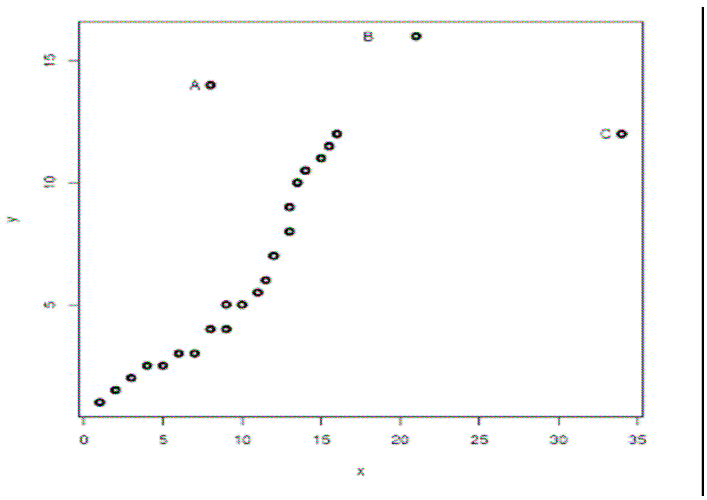


Fig. 3.3: Diversi tipi di dati anomali: (A) in y, (B) in x e y, (C) in x.

Una ampia classe di stimatori $(\tilde{\beta},\tilde{\sigma})$, di tipo M per parametri di scala e regressione è definita da

$$\sum_{i=1}^{n} \psi(y_i;\beta,\sigma) = \begin{pmatrix} \sum s(x_i)\psi_\beta\{r_i v(x_i)\}x_i \\ \sum \psi_\sigma(r_i) \end{pmatrix}, \qquad (3.10)$$

dove *s(.), v(.), ψ_β(.) e ψ_σ(.)* sono funzioni note appropriate (cfr. Hampel et al., 1986, cap.6). Per *s(x) = v(x) = 1* e $\psi_\beta(.) = \psi_k(.)$ si ottiene lo stimatore di Huber. Alternativamente, la scelta *s(x) = 1/v(x), v(x)* = ||x|| e $\psi_\beta(.) = \psi_k(.)$ definisce lo stimatore di Hampel-Krasker, che costituisce una soluzione al teorema di Hampel quando il modello di riferimento è quello normale (cfr. Hampel *et al*., 1986). Una scelta usuale per la funzione $\psi_\sigma(.)$, è $\psi_\sigma(\cdot) = \psi_{k_1}^2(\cdot) - \gamma(k_1)$, dove k_1 e $\gamma(k_1)$ sono delle costanti appropriate.

Dalla (3.10), è immediato ricavare le espressioni per la IF degli stimatori di β e σ e per la loro varianza asintotica. In particolare, come descritto nel paragrafo precedente, la funzione di influenza è data da

$$IF(c;\tilde{\beta},\tilde{\sigma}) = M_\psi^{-1}(\tilde{\beta},\tilde{\sigma})\psi(c;\tilde{\beta},\tilde{\sigma}),$$

mentre la varianza asintotica è

$$V_\psi(\beta,\sigma) = M_\psi^{-1}(\beta,\sigma)A_\psi(\beta,\sigma)M_\psi^{-T}(\beta,\sigma).$$

Pertanto, per *n* sufficientemente elevato, test e regioni di confidenza per (β, σ) possono essere costruiti nel modo usuale, utilizzando una stima della varianza asintotica $V_\psi(\beta,\sigma)$.

Infine si osserva che anche per il modello di regressione e scala con errori non normali è possibile definire lo stimatore ottimale di Hampel chiamato OBRE, ossia lo stimatore B-robusto a minima varianza. Più precisamente, l'OBRE è ottimale nel senso che costituisce lo stimatore di tipo M che minimizza la traccia della matrice di varianza asintotica sotto al vinvolo che la IF risulti limitata. Esistono varie versioni dell'OBRE, che dipendono dal modo in cui si decide di limitare la IF (cfr. Hampel *et al*., 1986). Il metodo più utilizzato (si veda anche Carroll e Ruppert, 1988) è quello basato sull'OBRE standardizzato. Dato un limite $k = c\sqrt{p}$ sulla IF, ove *p* indica il

numero totale dei parametri coinvolti nel modello, ossia la dimensione di $\theta = (\beta, \sigma)$, lo stimatore OBRE è definito implicitamente dall'equazione di stima

$$\sum_{i=1}^{n} \psi(y_i;\theta) = \sum_{i=1}^{n} \{\ell_*(\theta;y_i) - a(\theta)\}\, w_i(y_i;\theta) = 0 \ , \tag{3.11}$$

ove

$$w_i(y;\theta) = \min\left\{1, \frac{k}{||\ell_*(\theta;y) - a(\theta)||_{A(\theta)}}\right\} \ , \tag{3.12}$$

$||x||_{A(\theta)} = [x^{\mathrm{T}} A(\theta)^{-1} x]^{1/2}$ e la matrice $A(\theta)$ di dimensioni p x p e il vettore p-dimensionale $a(\theta)$ sono definiti tramite le condizioni

$$A(\theta) = \frac{1}{n} E[\psi(y;\theta)\psi(y;\theta)^{\mathrm{T}}] \quad \text{e} \quad E[\psi(y;\theta)] = 0 \ . \tag{3.13}$$

Per l'efficienza, l'OBRE deve essere il più simile possibile allo stimatore di massima verosimiglianza per i valori di y contenuti nella maggioranza dei dati, ossia per i valori non influenti di y. Di conseguenza, la sua funzione $\psi(.;\theta)$ è pari alla funzione score per questi valori delle osservazioni. Invece, poichè la IF dell'OBRE è proporzionale alla funzione $\psi(y;\theta)$, per ottenere una IF limitata, si deve troncare $l_*(\theta;y)$ quando il valore k viene superato. Questo è garantito dai pesi (3.12). Nelle applicazioni, questi pesi sono molto utili in quanto essi permettono di definire automaticamente le osservazioni che sono state considerate dall'OBRE più o meno distanti dalla maggioranza dei dati. In particolare, grazie a questi pesi, è possibile definire il livello di contaminazione nei dati. La matrice $A(\theta)$ e il vettore $a(\theta)$ possono essere interpretati come i moltiplicatori di Lagrange derivanti dai vincoli di una IF limitata e di Fisher consistenza. In particolare, il vettore $a(\theta)$ è necessario per assicurare che l'equazione di stima sia non distorta. Il parametro k è il limite sulla IF e può essere interpretato come il regolatore tra le richieste di robustezza e di efficienza: per k piccolo si guadagna robustezza ma si perde efficienza, e viceversa per k grande. Lo stimatore più robusto è ottenuto per $k = \sqrt{p}$. Al contrario, $k \to \infty$ produce lo stimatore di massima verosimiglianza (lo stimatore più efficiente, ma non

robusto). Nella pratica, la scelta di *k* dipende in generale dal modello in esame. Infine, si osservi che di fatto θ, $A(\theta)$, $a(\theta)$ e $w_i(y; \theta)$ devono essere determinati simultaneamente risolvendo le (3.11), (3.12) e (3.13). Per risolvere queste equazioni si può usare un algoritmo iterativo basato sul metodo di Newton-Raphson (cfr. Hampel et al., 1986, Carroll e Ruppert, 1988, Peracchi, 1990). Più precisamente, l'algoritmo può essere definito dai seguenti passi:

1. Si fissano un livello di precisione ε o il numero massimo di iterazioni *J* che saranno usati nell'algoritmo. Sia θ_0 un valore iniziale per il parametro Θ o una sua stima preliminare, ad esempio la stima di massima verosimiglianza. Si fissi $j = 1$, $w_i = 1$ per ogni *i* e $a(\theta_0) = 0$.

2. Si calcola:

$$A(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} w_i^2(y_i;\theta_0)(\ell_*(\theta_0;y_i) - a(\theta_0))(\ell_*(\theta_0;y_i) - a(\theta_0))^{\mathrm{T}}.$$

3. Usando la (3.12) si aggiornano i pesi come

$$w_i(y;\theta_0) = \min\{1, k/||\ell_*(\theta_0;y) - a(\theta_0)||_{A(\theta_0)}\}.$$

4. Usando questi pesi si risolve la seguente equazione in θ:

$$0 = \sum_{i=1}^{n} w_i(y_i;\theta_0)\{\ell_*(\tilde{\theta};y_i) - a(\theta_0)\}$$

5. Se $\max|(\tilde{\theta} - \theta_0)/\theta_0| < \varepsilon$ o $j \geq J$ allora ci si ferma. Altrimenti si aggiorna $a(\theta)$ come $a(\tilde{\theta}) = \sum_{i=1}^{n} \ell_*(\tilde{\theta};y_i) w_i(y_i;\theta_0) / \sum_{i=1}^{n} w_i(y_i;\theta_0)$ e si ritorna al passo 2. *Fine*

Questo algoritmo è convergente a condizione che il valore iniziale ϑ_0 sia vicino alla soluzione, ma la convergenza può essere lenta.

3.5 L'indice di determinazione R^2 robusto

E' piuttosto evidente che, nel modello lineare, in caso di errata specificazione del

modello o in presenza di *outliers*, non solo le stime di massima verosimiglianza ne risentono, ma anche l'indice di determinazione R^2, non solo attraverso i coefficienti stimati $\hat{ß}$ utilizzati per calcolare la variabile risposta, ma anche attraverso i residui o scarti $y_i - \bar{y}$. In ambito robusto, tra le tecniche più recenti per la validazione del modello stimato è stato proposto l'indice di determinazione R^2 robusto (*Renaud, 2009*), che è definito come

$$R_w^2 = \left(\frac{\sum_{i=1}^{n} w_i (y_i - \bar{y}_w)(\hat{y}_i - \bar{\hat{y}}_w)}{\sqrt{\sum_{i=1}^{n} w_i (y_i - \bar{y}_w)^2 \sum_{i=1}^{n} w_i (\hat{y}_i - \bar{\hat{y}}_w)^2}} \right)^2 , \qquad (3.14)$$

dove

$\bar{y}_w = (1/\sum wi)\sum w_i y_i$,

$\bar{\hat{y}}_w = (1/\sum wi)\sum w_i \hat{y}_i$ e

$$w_i = w(r_i; c) = \begin{cases} \left(\left(\frac{r_i}{c} \right)^2 - 1 \right)^2 & se \quad |r_i| \le c \\ 0 & se \quad |r_i| > c \end{cases}$$

I pesi w_i e i valori stimati $\hat{y}_i$ sono prodotti dalle stime di regressione robusta. Dalla (3.14) è possibile ricavare una formulazione più semplice dell' R_w^2 basata sulla somma totale dei quadrati, ossia

$$\tilde{R}_w^2 = \frac{\sum_{i=1}^{n} w_i (y_i - \bar{y}_w)^2 - \sum_{i=1}^{n} w_i (\hat{y}_i - \bar{\hat{y}}_i)^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y}_w)^2}. \qquad (3.15)$$

Si tratta di un indice normalizzato, compreso tra 0 e 1, quindi confrontabile anche con l'R^2 ottenuto stimando i coefficienti con un modello di regressione lineare, come

vedremo nel capitolo successivo.

3.6 Conclusioni

In questo capitolo è stata presentata una breve rassegna su alcune tecniche di regressione robusta e sulla valutazione della bontà del modello mediante l'indice di determinazione R^2 robusto. Per la semplicità di analisi e comprensione, la regressione robusta può essere considerata una valida alternativa alla regressione lineare quando non sono soddisfatte le ipotesi del modello e quando la stabilità rispetto alla specificazione scorretta o alla contaminazione è richiesta. Nel Capitolo successivo verrà applicata la regressione robusta ai dati aziendali, come modello alternativo alla regressione lineare vista nel Capitolo 2.

CAPITOLO 4

Applicazione della Regressione Robusta

Nel capitolo precedente, dedicato ad una rassegna sulla teoria della robustezza, con particolare attenzione al modello di regressione, abbiamo posto le basi, da un punto di vista teorico, per poter analizzare e confrontare i risultati cui giungiamo stimando i coefficienti con il nuovo modello di regressione robusta utilizzando lo stimatore di Huber. I comandi di ***R*** *per la regressione robusta sono raccolti nella libreria "MASS".*

4.1 Applicazione del modello ai dati

A differenza del modello di regressione lineare classico, introdotto nel Capitolo 2, nel modello di regressione robusta viene assegnato un peso a ciascuna osservazione dando meno importanza alle osservazioni anomale. Il metodo di stima robusto di seguito utilizzato è quello introdotto da Huber (1964). Partendo dal modello di regressione (3.8) visto nel Paragrafo 3.4, le stime dei coefficienti β sono ottenute come soluzione dell'equazione di stima

$$\sum_{i=1}^{n} \psi_k (r_i) x_i = 0 \Big|.$$

Quindi, utilizzando lo stesso procedimento visto per la regressione lineare (si veda Capitolo 2), il primo passo è calcolare le stime dei coefficienti β (si veda Tab 4.1) con un modello di regressione robusta sui dati della stagione precedente (072 – A/I 2007-2008).

		Coefficienti Robusti		
week_ord	week	ß1r_PO	ß2r_TG	ß3r_RT
1	39	4,9	52,41	2254,23
2	40	6,3	31,72	1495,9
3	41	5,07	29,4	660,49
4	42	6,49	13,34	516,2
5	43	6,49	9,06	362,65
6	44	6,25	8,5	285,94
7	45	6,36	9,04	141,8
8	46	6,22	8,61	115,41
9	47	6,53	9,48	59,28
10	48	4,63	7,3	98,1
11	49	4,26	6,26	104,02
12	50	3,26	7,26	103,66
13	51	2,63	6,27	104,85
14	52	2,66	6,25	72
15	01	2,66	6,25	72
16	02	2,38	5,92	66,15
17	03	2,06	5,59	67,12
18	04	1,98	7,89	-7,71
19	05	1,63	5,76	23,37
20	06	1,56	4,28	29,49
21	07	1,38	3,37	43,53
22	08	1,37	1,67	42,51
23	09	1,16	1,44	44,32
24	10	1,07	1,95	14,18
25	11	1,04	1,58	9,15
26	12	1,03	1,33	6,12
27	13	1,03	0,88	3,31
28	14	1,04	0,37	1,68
29	15	1,04	0,34	-0,58
30	16	1,02	0,38	0,09
31	17	1,02	0,11	2,08
32	18	1,01	0,14	0,59
33	19	1	0,24	-0,59
34	20	1	0,15	0,61
35	21	1	0,05	0,12
36	22	1	0,02	0,01
37	23	1	0,03	0,09
38	24	1	0,04	-0,28
39	25	1	0,02	0,09
40	26	1	0,005	0,033

Tab. 4.1: Coefficienti derivanti dalla regressione robusta sulla stagione precedente (p-value < 0.05).

Rappresentando graficamente i risultati della Tab. 4.1, nelle Fig. 4.1-4.3 si nota un andamento decrescente col passare delle settimane; infatti, trattandosi di dati progressivi che aumentano di settimana in settimana, con una variabile dipendente che rappresenta il totale dei metri venduti nella stagione, è normale che nelle prime settimane il valore delle stime dei β sia alto e che diminuisca col passare delle settimane.

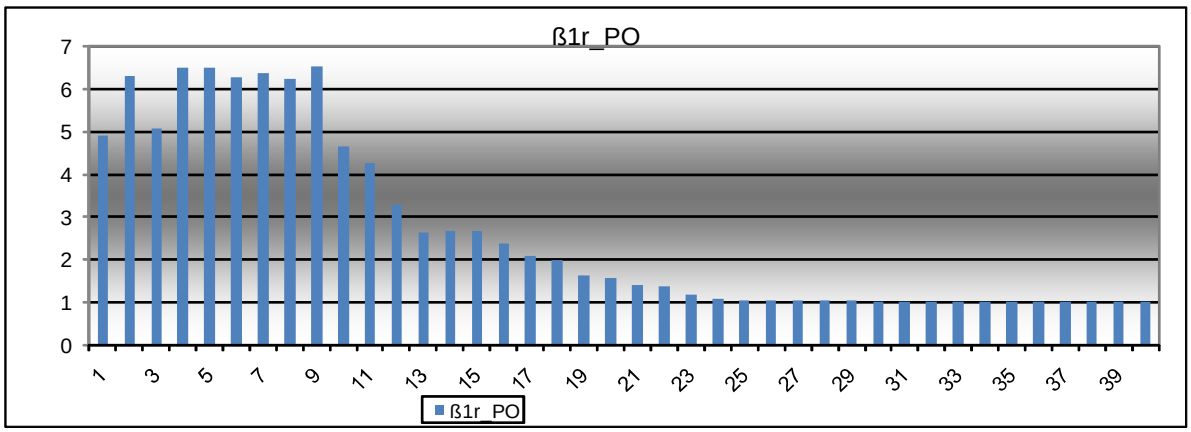


Figura 4.1: Coeff. per la variabile PO.

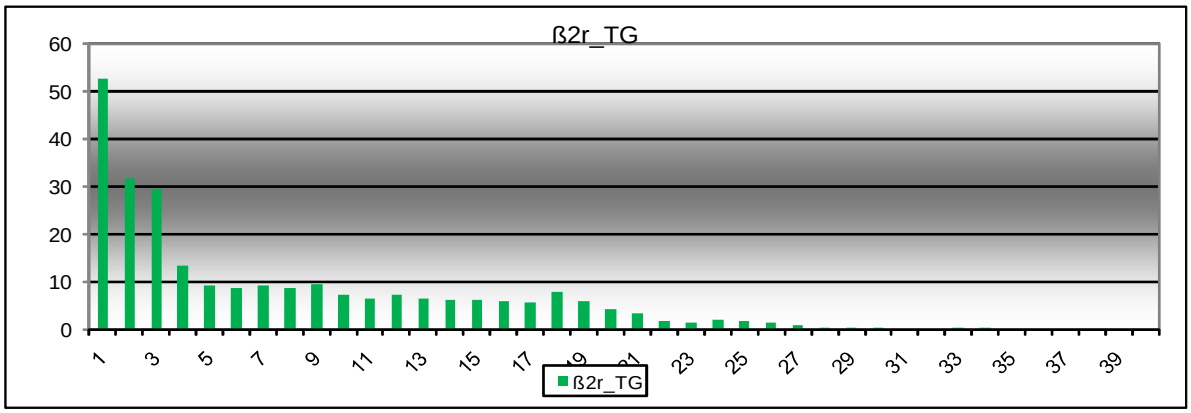


Figura 4.2: Coeff. per la variabile TG.

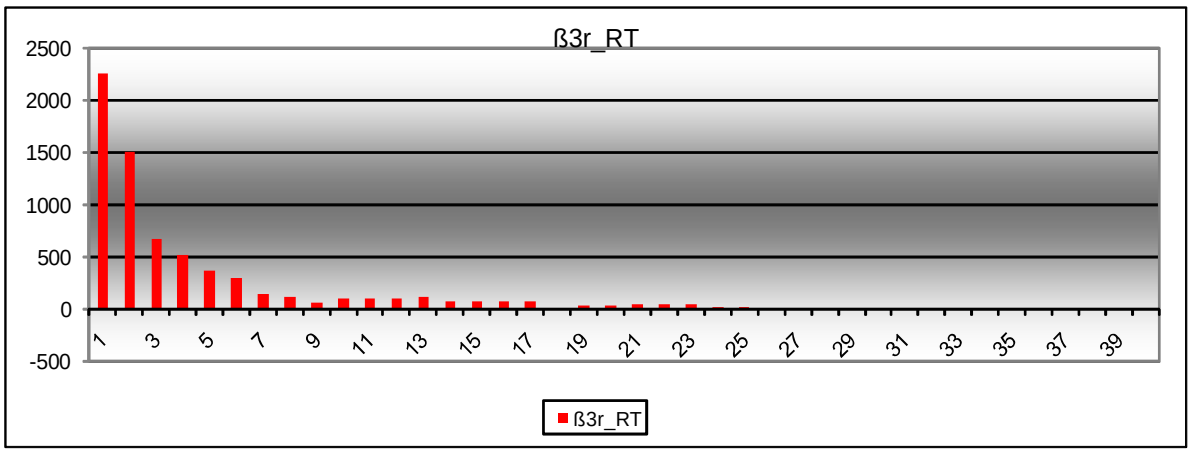


Figura 4.3: Coeff. per la variabile TG.

Nonostante la significatività dei coefficienti, si nota che il contributo della variabile RT è minimo rispetto alle altre due variabili, soprattutto rispetto alla variabile PO che rappresenta gli ordini di produzione veri e propri. Se poniamo a confronto le stime ottenute con il nuovo modello di regressione robusta rispetto a quelle ottenute con il

modello di regressione lineare (cfr. Paragrafo 2.2, Tab. 2.3 e Fig. 2.1-2.3) si nota un miglioramento delle stime col passaggio a tecniche robuste: non vi sono più stime di valore negativo come nel modello lineare che ha portato a vincolare i coefficienti stimati per evitare di ottenere previsioni minori di quanto già effettivamente ordinato.

Il secondo passo, dopo aver calcolato le stime dei coefficienti β, è calcolare la previsione totale dei metri che verranno venduti. A tal fine sono state applicate le stime robuste alle variabili PO, TG ed RT sui dati della stagione corrente (082 – 2008-2009) per ogni settimana.

Nella Tab. 4.2 sono stati aggregati i dati per settimana (trascurando per ora la famiglia) e sono messe a confronto le previsioni con i consuntivi, calcolando il relativo errore di previsione così come è stato fatto per il modello precedente visto nel Capitolo 2 (si veda Tabella 2.4). Sono state riportate in dettaglio le seguenti voci:

- il campo “week_ord” è stato inserito per comodità e ordina in senso crescente le settimane;
- il campo “DTOT_robust” rappresenta la previsione calcolata con i coefficienti derivanti dal modello di regressione robusta;
- il campo “DTOT_Act” rappresenta il consuntivo a fine stagione (ignoto in fase di previsione). Il dato è costante per tutte le settimane perché l’obiettivo è quello di “*prevedere quanti metri entreranno nella stagione in corso*”: quindi ogni previsione andrà confrontata con questo valore;
- il campo “Delta Robust vs Act” rappresenta lo scostamento calcolato come differenza tra la previsione e il consuntivo;
- il campo “Errore Rob vs Act%” rappresenta invece l’errore percentuale calcolato come rapporto tra delta e consuntivo;
- i campi “Errore Puro vs Act%” e “Errore Vinc vs Act%” rappresentano l’errore percentuale derivante dalla previsione con il modello di regressione lineare (si veda Tab. 2.4).

StCom: Collezione & Esclusivo			Prev_robusta	Consuntivo a fine stg.	Scostamento vs consuntivo	Regressione Robusta		Regressione Lineare	
Anno	week	week_ord	DTOT_robust	DTOT_Act	Delta Robust vs Act	Errore Rob vs Act%		Errore Puro vs Act%	Errore Vinc vs Act%
2007	39	01	2.806.033	4.827.289	- 2.021.256	-41,9%		-24,9%	-20,6%
	40	02	2.583.986	4.827.289	- 2.243.303	-46,5%		-21,8%	-21,8%
	41	03	3.059.517	4.827.289	- 1.767.771	-36,6%		-13,6%	-9,6%
	42	04	2.925.353	4.827.289	- 1.901.936	-39,4%		-17,9%	-17,9%
	43	05	3.106.743	4.827.289	- 1.720.546	-35,6%		-9,7%	-9,7%
	44	06	3.510.577	4.827.289	- 1.316.712	-27,3%		-1,8%	-1,8%
	45	07	3.284.181	4.827.289	- 1.543.107	-32,0%		-6,5%	-6,5%
	46	08	3.872.669	4.827.289	- 954.620	-19,8%		10,8%	10,8%
	47	09	4.337.875	4.827.289	- 489.414	-10,1%		16,8%	16,8%
	48	10	4.333.580	4.827.289	- 493.709	-10,2%		25,6%	25,6%
	49	11	4.412.229	4.827.289	- 415.060	-8,6%		28,3%	28,3%
	50	12	4.319.646	4.827.289	- 507.643	-10,5%		20,1%	20,1%
	51	13	4.481.678	4.827.289	- 345.611	-7,2%		22,4%	22,4%
	52	14	4.371.542	4.827.289	- 455.747	-9,4%		19,1%	19,1%
2008	01	15	4.490.971	4.827.289	- 336.318	-7,0%		22,3%	22,3%
	02	16	4.589.923	4.827.289	- 237.366	-4,9%		24,7%	24,7%
	03	17	4.357.276	4.827.289	- 470.013	-9,7%		23,4%	28,6%
	04	18	4.639.528	4.827.289	- 187.761	-3,9%		32,5%	62,3%
	05	19	4.307.208	4.827.289	- 520.081	-10,8%		1,9%	5,0%
	06	20	4.796.025	4.827.289	- 31.264	-0,6%		11,9%	13,7%
	07	21	4.739.353	4.827.289	- 87.936	-1,8%		13,9%	17,6%
	08	22	4.881.969	4.827.289	54.681	1,1%		16,3%	20,7%
	09	23	4.543.997	4.827.289	- 283.292	-5,9%		12,6%	18,6%
	10	24	4.334.093	4.827.289	- 493.195	-10,2%		10,5%	24,5%
	11	25	4.363.585	4.827.289	- 463.704	-9,6%		12,1%	24,7%
	12	26	4.399.987	4.827.289	- 427.301	-8,9%		12,4%	24,8%
	13	27	4.399.117	4.827.289	- 428.172	-8,9%		10,9%	22,2%
	14	28	4.470.436	4.827.289	- 356.853	-7,4%		13,0%	25,2%
	15	29	4.651.974	4.827.289	- 175.315	-3,6%		17,4%	30,2%
	16	30	4.660.192	4.827.289	- 167.096	-3,5%		17,8%	30,9%
	17	31	4.625.863	4.827.289	- 201.426	-4,2%		16,5%	30,5%
	18	32	4.659.832	4.827.289	- 167.457	-3,5%		17,2%	31,7%
	19	33	4.701.797	4.827.289	- 125.492	-2,6%		18,1%	32,5%
	20	34	4.735.912	4.827.289	- 91.377	-1,9%		18,7%	32,9%
	21	35	4.799.872	4.827.289	- 27.417	-0,6%		20,8%	35,2%
	22	36	4.809.658	4.827.289	- 17.631	-0,4%		1,5%	2,6%
	23	37	4.804.781	4.827.289	- 22.508	-0,5%		1,4%	2,4%
	24	38	4.814.563	4.827.289	- 12.726	-0,3%		1,5%	1,9%
	25	39	4.822.889	4.827.289	- 4.400	-0,1%		0,8%	1,5%
	26	40	4.828.939	4.827.289	1.650	0,0%		0,8%	1,3%

Tab. 4.2: Analisi scostamenti di previsione vs consuntivo.

Ad eccezione delle prime settimane in cui si ottiene un errore di previsione piuttosto alto, dalla settimana 09 in poi (week_ord = 09) l'errore non supera il 10% : un risultato più che soddisfacente se confrontato con l'errore di previsione ottenuto dal modello precedente. Nel paragrafo seguente verranno messi a confronto i due modelli

sulla base dell'errore di previsione ottenuto.

4.2 Confronto con il modello precedente

L'utilizzo della regressione robusta rende più omogenee le previsioni settimanali (si veda Tab. 4.2) perché tiene conto di vari allontanamenti dalle ipotesi del modello classico, a differenza della regressione lineare (si veda Tab. 2.4) con la quale si ottengono previsioni irregolari che discostano molto dalla realtà. Nella Fig. 4.4 si pone a confronto l'errore di previsione settimanale ottenuto con i due modelli:

- La serie "Errore Puro vs Act %" rappresenta l'errore percentuale calcolato utilizzando le stime dei coefficienti ottenute da un modello di regressione lineare semplice.
- La serie "Errore Vinc vs Act %" rappresenta l'errore percentuale calcolato utilizzando le stime vincolate (si veda Paragrafo 2.2) dei coefficienti ottenute da un modello di regressione lineare semplice.
- La serie "Errore Robust vs Act %" rappresenta l'errore percentuale calcolato utilizzando le stime dei coefficienti ottenute da un modello di regressione robusta.

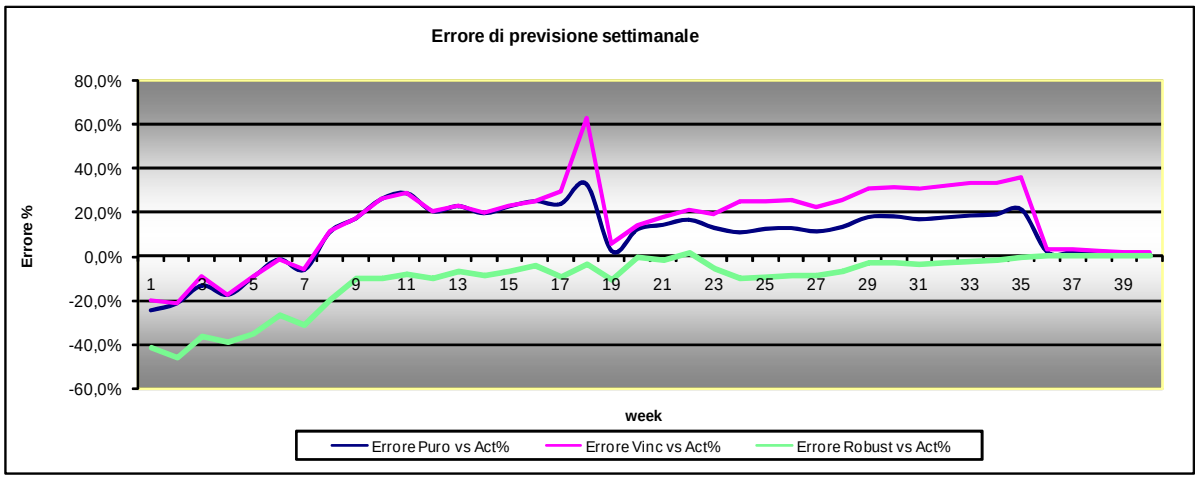


Fig. 4.4: Errore di previsione settimanale sia per famiglie di collezione che esclusivi.

Dalla Fig. 4.4 si nota un miglioramento delle previsioni mediante l'utilizzo di un modello di regressione robusta: a differenza del modello lineare che tende a sovrastimare le previsioni nelle settimane centrali, addirittura con un picco nella settimana 18, con la regressione robusta l'errore è sempre negativo (le previsioni non sono sovrastimate). Questo è un aspetto importante perché sovrastimare le vendite vuol dire aumentare le scorte di prodotti finiti a magazzino. Un'operazione, quest'ultima dannosa per l'azienda in quanto il costo di mantenimento del magazzino può salire fino al 30-40% del valore dei prodotti a stock, o addirittura rischiosa, quando il capitale immobilizzato risulti particolarmente elevato e le scorte siano soggette ad obsolescenza, ovvero a svalutazione economica.

Anche per le sole famiglie di collezione (Fig. 4.5) l'utilizzo della regressione robusta sembra produrre buoni risultati, a differenza della regressione lineare: l'errore di previsione è nettamente migliore e prossimo allo zero.

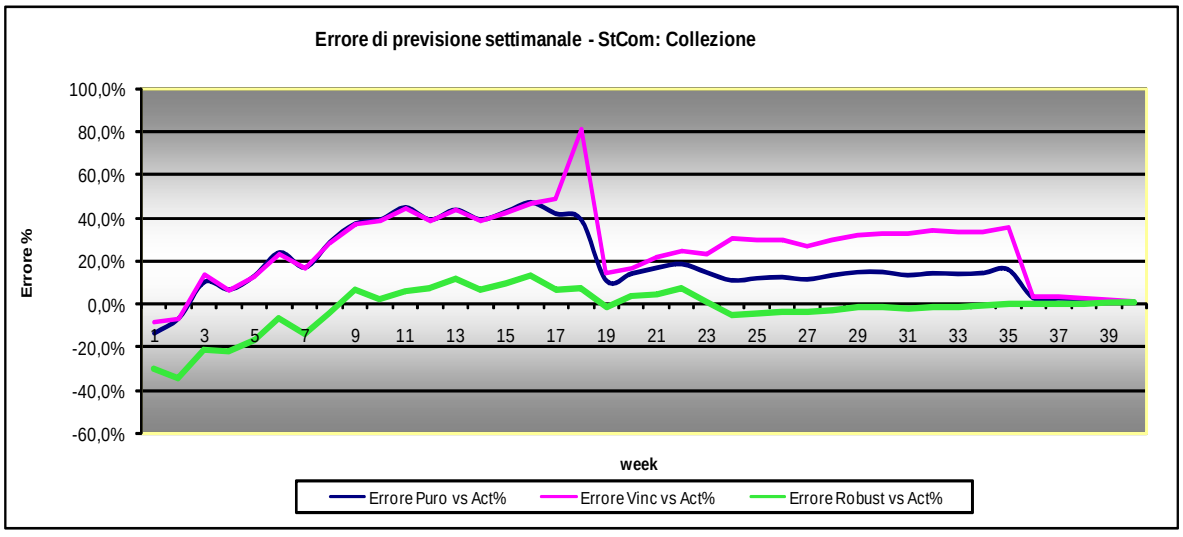


Fig. 4.5: Errore di previsione settimanale per famiglie di collezione.

4.3 Valutazione del modello: l'R^2 robusto

Tra le tecniche più recenti per la validazione del modello stimato in ambito robusto vi è l'indice di determinazione R^2 robusto (si veda il Paragrafo 3.5). La Fig. 4.6 mostra l'indice calcolato per ogni settimana sul modello di regressione robusta: si nota come già dalle prime settimane il modello stimato si adatta bene ai dati (superando l80%) e l'indice segue un andamento crescente, sempre più vicino al suo valore massimo (100%).

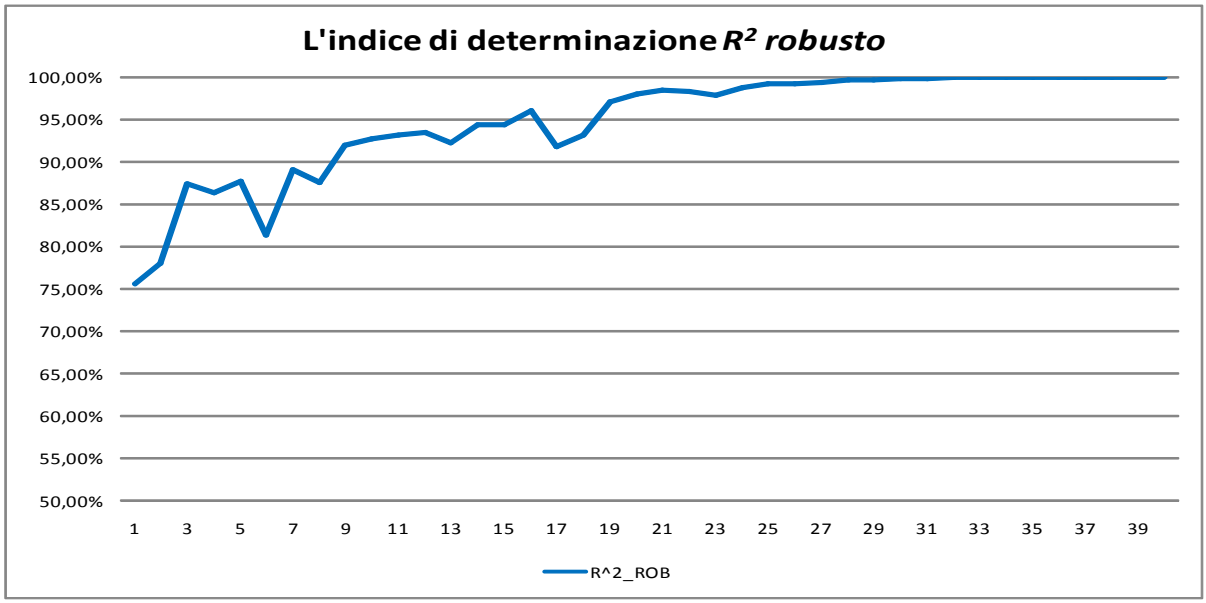


Fig. 4.6: Andamento settimanale dell'indice di determinazione R^2 robusto.

Il passaggio a un modello di regressione robusta porta, quindi, ad una stima più accurata del modello anche in termini di R^2. Infatti, confrontando i due indici in Fig. 4.7 si nota un netto miglioramento con il passaggio a tecniche robuste (R^2_LIN si riferisce all'indice basato sul modello di regressione lineare, mentre R^2_ROB si

riferisce a quello basato sul modello di regressione robusta).

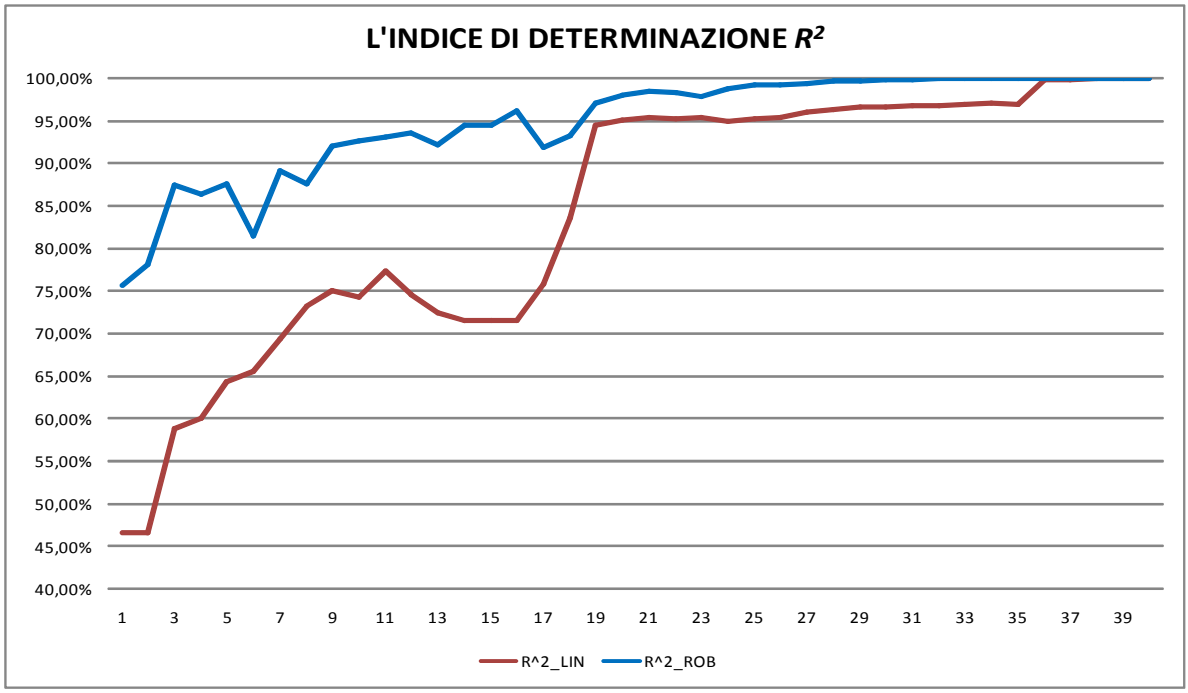


Figura 4.7: Indici di determinazione a confronto.

4.4 Presentazione dei comandi in R

Di seguito sono richiamati i comandi di R per la regressione robusta per i dati esposti nel Capitolo 2. La procedura per giungere alla previsione resta la stessa, ciò che cambia è il metodo di stima. Quindi, innanzitutto si caricano i dati per ogni settimana relativi alla stagione passata con il comando

```
> prec <-read.csv2("C:/inputdati/072/01_072.csv",
header=TRUE, sep=";", quote="\"", dec=",", row.names=1,
fill=TRUE)
> attach(prec)
```

In questo modo è stato caricato il file 01_072.csv, dove 01 indica la settimana che si sta caricando e 072 la stagione passata.

Il secondo passo è stimare i dati con la regressione robusta. Il comando che usa che usa la funzione di Huber è `rlm()`(*Robust Linear Model*) contenuto nella libreria `MASS` :

```
> library(MASS)
> fitprec<- rlm(DTOT~ PO+TG+RT-1, psi=psi.huber)
```

Questo comando permette di ottenere le stime dei coefficienti di regressione robusta delle variabili esplicative che servono per la stima del modello sui dati della stagione corrente; infatti, il passo successivo è quello di caricare i dati della stagione che si vuole prevedere (si veda anche Paragrafo 2.5):

```
> act <- read.csv2("C:/input dati/082/01_082.csv",
header=TRUE, sep=";", quote="\"", dec=",", fill=TRUE)
> attach(act)
```

Ora con il comando *predict* calcolo la previsione, utilizzando i coefficienti di regressione robusti:

```
> dtot_01 <- data.frame(predict(fitprec,
newdata=data.frame(PO=act[,2], TG=act[,3], RT=act[,4]),
se.fit=T, , interval= “confidence”), row.names=act[,1])
```

Le previsioni così ottenute sono state esportate in Excel con il comando *write.table* :

```
>write.table(dtot_01,file="C:/input dati/ACT/01_act.xls",
append=FALSE, quote=TRUE, sep="\t", dec=",",
row.names=TRUE).
```

La procedura appena presentata per il calcolo della previsione alla settimana 01, è speculare per tutte le altre settimane.

In riguardo al calcolo di R^2 robusto, trattandosi di una tecnica recente per la validazione del modello, nel programma statistico R non è stata ancora implementata

una funzione per calcolare questo indice. Di conseguenza, il calcolo è stato effettuato con Microsoft Excel ed esportando da R solo i pesi (`fitprec$w`) e i valori stimati (`fitprec$fit`) per ogni settimana con il comando *write.table*:

```
> PesiW <- fitprec$w
```

```
>write.table(PesiW,file="C:/inputdati/PESI_W/PesiW.xls",
append=FALSE, quote=TRUE, sep="\t", dec=",",
row.names=TRUE)
```

```
>write.table(fitprec40$fit,file="C:/inputdati/FITPREC/
fitprec.xls", append=FALSE, quote=TRUE, sep="\t",
dec=",", row.names=TRUE)
```

4.5 Conclusioni

A completamento di quanto esposto nel Capitolo 3, in questo capitolo è stata presentata l'applicazione del nuovo modello che verrà adottato dall'azienda a scopo previsivo. Si è scelto un modello di regressione robusta per venire incontro alle esigenze dell'azienda, la quale chiedeva un modello semplice, basato comunque sulle tecniche di regressione e che stimasse in modo più opportuno la variabile d'interesse.

Infatti, si è visto che il nuovo modello porta a dei risultati soddisfacenti, con errori di previsione molto più bassi rispetto al modello di regressione lineare (si veda Tab. 4.2). Anche dall'interpretazione dell'indice di determinazione R^2 si desume che, utilizzando una tecnica robusta, il modello si adatta meglio ai dati.

CONCLUSIONI

Quando un'azienda non adotta un corretto sistema di previsione, o lo applica in maniera incompleta, la naturale conseguenza è l'adattamento del livello di scorte di prodotti finiti alla variabilità della domanda. Una migliore previsione della domanda consente di migliorare la propria capacità di risposta al mercato, contenendo allo stesso tempo i costi, permettendo di pianificare meglio la produzione e riducendo le scorte di prodotti finiti. In secondo luogo, una migliore previsione consente di effettuare acquisti e gestire le scorte di materie prime in modo migliore, aumentando la disponibilità e riducendo gli obsoleti. Per migliorare l'economicità della gestione e per la realizzazione di un management efficiente, le previsioni sono quindi un elemento primario. Questa tesi discute alcuni aspetti di un'analisi di regressione per la previsione delle vendite al caso aziendale del Cotonificio Albini SPA che progetta e produce, a catalogo e su commessa, tessuti per camiceria. Obiettivo principale del Cotonificio Albini è essenzialmente rimanere leader del mercato nelle fasce di prodotti di media-alta qualità. E' per questo che, da qualche anno, l'azienda ha implementato un sistema di previsione della domanda per meglio pianificare gli acquisti di materie prime, verificare l'effettiva capacità produttiva e garantire così ai propri clienti un livello di servizio migliore e tempi di consegna più brevi. La tesi prevede lo studio critico del modello di regressione attualmente utilizzato dall'azienda per il calcolo della previsione e la proposta di un nuovo modello semplice, sempre nell'ambito della regressione lineare, che migliori quello attualmente utilizzato. Innanzitutto è stato esposto il metodo di previsione utilizzato attualmente dall'azienda per stimare la quantità di tessuto (metri) che verrà ordinata in una stagione, utilizzando un modello di regressione lineare, basato sulle stime di massima verosimiglianza, opportunamente vincolate per evitare di ottenere previsioni negative o inferiori rispetto a quanto già ordinato. Dal confronto tra le previsioni e i

dati reali è emerso che il modello produce previsioni sovrastimate. Neanche il passaggio alle stime vincolate produce effetti migliorativi. La causa dell'errata specificazione del modello è legata alla violazione degli assunti di base del modello di regressione lineare. Al fine di venire incontro alle esigenze dell'azienda, la quale chiedeva un modello semplice, basato comunque su tecniche di regressione e che stimasse in modo più opportuno i dati, si è deciso di ricorrere alla regressione robusta. Essa costituisce un approccio alla statistica, in quanto ha lo scopo di salvaguardare rispetto a eventuali deviazioni delle ipotesi statistiche assunte. Quando è richiesta là stabilità rispetto alla specificazione scorretta oppure in presenza di elevati valori anomali (*autliers*), la regressione robusta può essere considerata una valida alternativa alla regressione lineare, per la semplicità di analisi e comprensione. Infatti, l'utilizzo della regressione robusta rende più omogenee le previsioni settimanali perché tiene conto di vari allontanamenti dalle ipotesi del modello classico, a differenza della regressione lineare con la quale si ottengono previsioni irregolari che discostano molto dalla realtà. Nella Fig. 1 si pone a confronto l'errore di previsione settimanale ottenuto con i due modelli:

- La serie "Errore Puro vs Act %" rappresenta l'errore percentuale calcolato utilizzando le stime dei coefficienti ottenute da un modello di regressione lineare semplice.
- La serie "Errore Vinc vs Act %" rappresenta l'errore percentuale calcolato utilizzando le stime dei coefficienti vincolate (a zero nel caso di coefficienti negativi), per non ottenere previsioni minori rispetto a quanto già effettivamente ordinato.
- La serie "Errore Robust vs Act %" rappresenta l'errore percentuale calcolato utilizzando le stime dei coefficienti ottenute da un modello di regressione robusta.

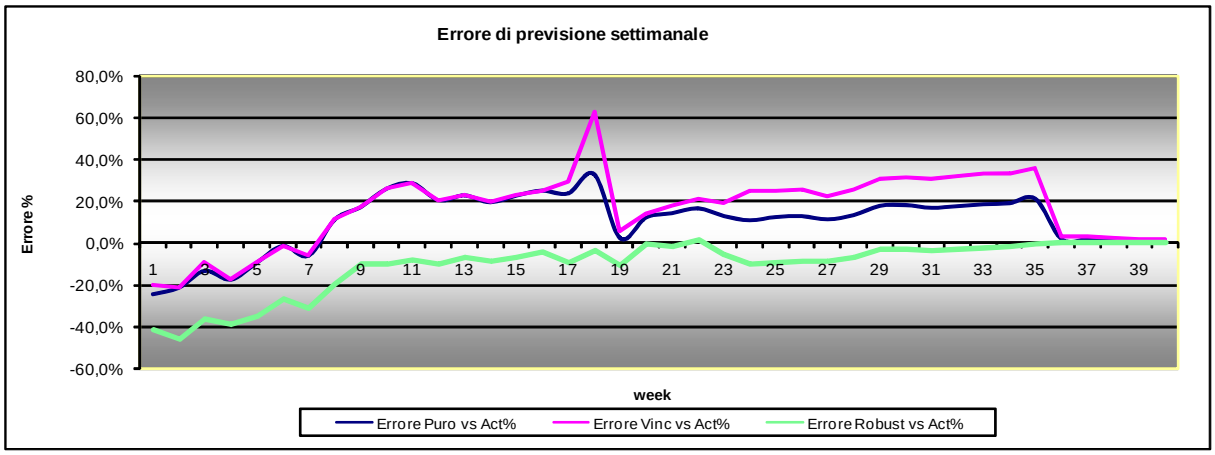


Fig. 1: Errore di previsione settimanale sia per famiglie di collezione che esclusivi.

Dalla Fig. 1 si nota un miglioramento delle previsioni mediante l'utilizzo di un modello di regressione robusta: a differenza del modello lineare che tende a sovrastimare le previsioni nelle settimane centrali, addirittura con un picco nella settimana 18, con la regressione robusta l'errore è sempre negativo (le previsioni non sono sovrastimate). Questo è un aspetto importante perché sovrastimare le vendite vuol dire aumentare le scorte di prodotti finiti a magazzino. Un'operazione, quest'ultima dannosa per l'azienda in quanto il costo di mantenimento del magazzino può salire fino al 30-40% del valore dei prodotti a stock, o addirittura rischiosa, quando il capitale immobilizzato risulti particolarmente elevato e le scorte siano soggette ad obsolescenza, ovvero a svalutazione economica.

Infine, per valutare la bontà del modello adottato, è stato calcolato il coefficiente di determinazione R^2. Confrontando i due indici in Fig. 2 si nota un netto miglioramento con il passaggio a tecniche robuste (R2_LIN si riferisce all'indice basato sul modello di regressione lineare, mentre R2_ROB si riferisce a quello basato sul modello di regressione robusta). Si nota come l'R^2 robusto si adatta bene ai dati già dalle prime settimane, seguendo un andamento crescente, sempre più vicino al suo valore massimo (100%).

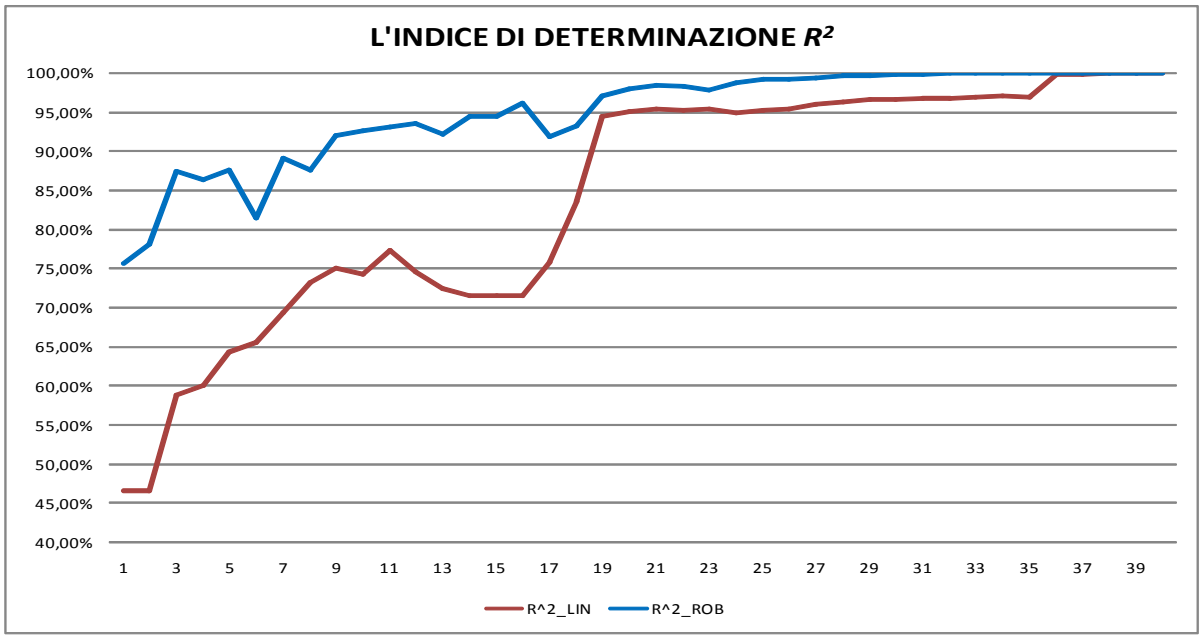


Fig. 2: Indici di determinazione a confronto.

Complessivamente si può affermare che il nuovo modello ha portato a dei risultati abbastanza sensati, seppur preliminari. Si tratta comunque di una prima applicazione, volta a migliorare il modello attualmente utilizzato dall'azienda, con il vincolo di continuare ad utilizzare un modello di regressione.

BIBLIOGRAFIA

- Azzalini, A. (2000). *Inferenza statistica. Una presentazione basata sul concetto di verosimiglianza.* Springer.
- Bortot P., Ventura L., Salvan A., (2000). *Inferenza statistica: Applicazioni con S-PLUS e R*. Cedam, Padova
- Carroll, R.J., Ruppert D. (1988) . *Transformation and Weighting in Regression.* Chapman and Hall, New York.
- Castelli, M. (2007). *Campioni del Mondo. Storie di Uomini, storie di Imprese.* Il Sole 24 Ore, Milano.
- Desmond, A.F. (1997). Optima Estimating functions, quasi-likelihood and statistical modeling. *Journal of statistical planning and inference*.
- Fox, J. (2002). *An R and S-PLUS Companion to Applied Regression*. Sage Publications.
- Godambe,V.P. (1960). An optimum Property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics*.
- Hampel, F,R., Marazzi, A., Ronchetti, E., Rousseeuw, P.J., Stahel, W.A., Welsh, R.E. (1982). Handouts for the instructional meeting on robust statistical methods. *15th European Meeting of Statisticians*, Palermo, Italy.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* J. Wiley, New York.

- Helland, I. S. (1987). *On the interpretation and use of* R^2 *in regression analysis*. Biometrics.
- Huber, P.J. (1964). Robust Estimation of a Location Parameter. *The annals of mathematical statistics.*
- Huber, P.J. (1977). *Robust Statistical Procedures.* Regional Conference Series in Applied Mathematics, No. 27, Soc. Industry. Appl. Math, Philadelphia, Penn.
- Huber, P.J. (1981). *Robust Statistics*. J. Wiley, New York.
- Iacus S.M., Masarotto G. (2007). *Laboratorio di statistica con R,* McGraw-Hill.
- Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). Robust Statistics: Theory and Methods. Wiley, London
- Pace, L., Salvan, A. (2001). *Introduzione alla Statistica II. Inferenza, verosimiglianza, modelli.* Cedam, Padova.
- Pace, L., Salvan, A. (1996). *Teoria della Statistica. Metodi, modelli e approssimazioni asintotiche.* Cedam, Padova.
- Peracchi F. (1990). Bounded-influence estimators for the Tobit model, *Journal of Econometrics*.
- Rand R., Wilcox.(1997). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press.
- Rawlings J. O. (1988). *Applied regression Analysis*. Wadsworth.

- Renaud, O. (2009). A robust coefficient of determination for regression. M*ethodology and Data Analysis, Psichology Department. University of Geneva*.
- Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*.
- Tukey, J.W. (1970). *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass.
- Yohai, V.J., Stahel, W. A., and Zamar, R. H. (1991). *A procedure for robust estimation and inference in linear regression.* Spinger- Verlag.