UNIVERSITÀ DI PISA

CORSO DI STATISTICA:
ANALISI DELLE COMPONENTI PRINCIPALI E REGRESSIONE LINEARE

COLLEZIONE DI PROGETTI "*PISAFLIX*"

ANNO ACCADEMICO 2019-2020

STEFANO PETROCCHI

SOMMARIO

Introduzione 3
Scopo dell'Analisi 3
Dataset 3
Dataset 3
Fattori 3
Analisi 4
Analisi delle Correlazioni 4
Grafici dei Dataset 4
Grafici della Correlazione 4
Osservazioni 4
Analisi delle componenti principali 6
Scopo dell'Analisi 6
Grafici della Varianza 6
Composizione delle Componenti 7
Biplot 7
Conclusione 8
Regressione Lineare 8
Scopo dell'Analisi 8
Analisi Modello non Ottimizzato 8
Analisi Modello Ottimizzato 8
Analisi Modello Semplificato 9
Confronto tra Modelli 9
Analisi dei Residui 9
Conclusione 10

INTRODUZIONE

SCOPO DELL'ANALISI

Si ipotizza che l'analisi sia stata commissionata da una grossa società di sale cinematografiche "PisaFlix", con sedi distribuite in tutto il mondo. PisaFlix è interessata a massimizzare i profitti proiettando nelle proprie sale le pellicole per cui gli incassi mondiali previsti risultino essere maggiori.

Pertanto, gli scopi dell'analisi sono:

- Determinare i fattori principali che caratterizzano un film di successo e la loro influenza sugli incassi mondiali.
- Ottenere una buona previsione di quelli che possono essere gli incassi globali complessivi del film in modo da stimare i profitti che se ne possono ricavare e preferire la proiezione di quelli a maggior redditività.

DATASET

DATASET

L'analisi è stata svolta su due dataset (*ctrl + click* per aprire i link):

- Il primo costituito da *3738* osservazioni e *11* fattori numerici (filmEsteso.csv).
- Il secondo costituito da *326* osservazioni campionate senza rimpiazzo dal dataset precedente e *17* fattori numerici tra cui gli 11 del dataset precedente (film.csv).

I dataset sono frutto di *fuzzy (con fattore 1) inner join senza duplicati* effettuati tramite la funzione *Combina Query* della sezione *Dati* di *Excel* tra varie tabelle i cui link di riferimento sono allegati in Link.txt.

Si è preferito l'utilizzo di due dataset in quanto alcuni fattori, presenti nel secondo dataset, erano reperibili per un limitato numero di osservazioni. Il primo dataset serve a verificare la qualità del campionamento e che non siano state selezionate osservazioni di natura particolare. Ogni osservazione è riferita ad un singolo film uscito nella sale cinematografiche internazionali.

FATTORI

L'analisi è stata svolta sui seguenti fattori numerici:

- **Anno:** anno di pubblicazione del film (intero positivo).
- **Durata:** durata, espressa in minuti, del film (intero positivo).
- **Rating:** voto medio espresso sul sito IMDB da critica e utenti in una scala da 0 a 100 per il film (intero positivo).
- **Votanti:** numero di voti attraverso il quale è stato calcolato il rating del film (intero positivo).
- **LikesFilm:** numero di "likes" della pagina Facebook ufficiale o della "fanpage" con più likes del film (intero positivo).
- **LikesRegista:** numero di "likes" della pagina Facebook ufficiale o della "fanpage" con più likes del regista del film (intero positivo).
- **LikesProtagonista:** numero di "likes" della pagina Facebook ufficiale o della "fanpage" con più likes del protagonista del film (intero positivo).
- **Budget:** budget, espresso in dollari americani, ricavato da The Numbers, del film (intero positivo).
- **IncassoDomestico:** ricavi, espressi in dollari americani, nel paese d'origine del film da parte dello stesso, prelevati da The Numbers (intero positivo).
- **IncassoMondiale:** ricavi, espressi in dollari americani, ottenuti in tutto il mondo dal film, prelevati da The Numbers (intero positivo).
- **LunghezzaTrama:** numero di parole contenute della trama del film, prelevata da Wikipedia (intero positivo).
- **NumeroAttori:** numero di attori principali del film (intero positivo).
- **Morti:** numero di corpi senza vita preseti nel film, ricavati da Movie Body Counts (intero positivo).
- **SRR:** Sexual References Ratio è il numero di parole più comini riguardanti gli schemi sessuali (selezionate in questa ricerca) contenute nella trama del film, divise per la lunghezza del film, moltiplicate per 100 in modo da ottenerne la percentuale (decimale positivo).
- **WikiPageQuality:** qualità della pagina Wikipedia del film, prelevata da wikirank (decimale positivo).
- **BestWikiRankLocale:** posizione massima storica, della pagina del film, nel ranking delle pagine Wikipedia della nazione d'origine, prelevata da wikirank (intero positivo, valore massimo *9999°*).
- **BestWikiRankGlobale:** posizione massima storica globale nel ranking di Wikipedia della pagina del film, prelevata da wikirank (intero positivo, valore massimo *9999°*).

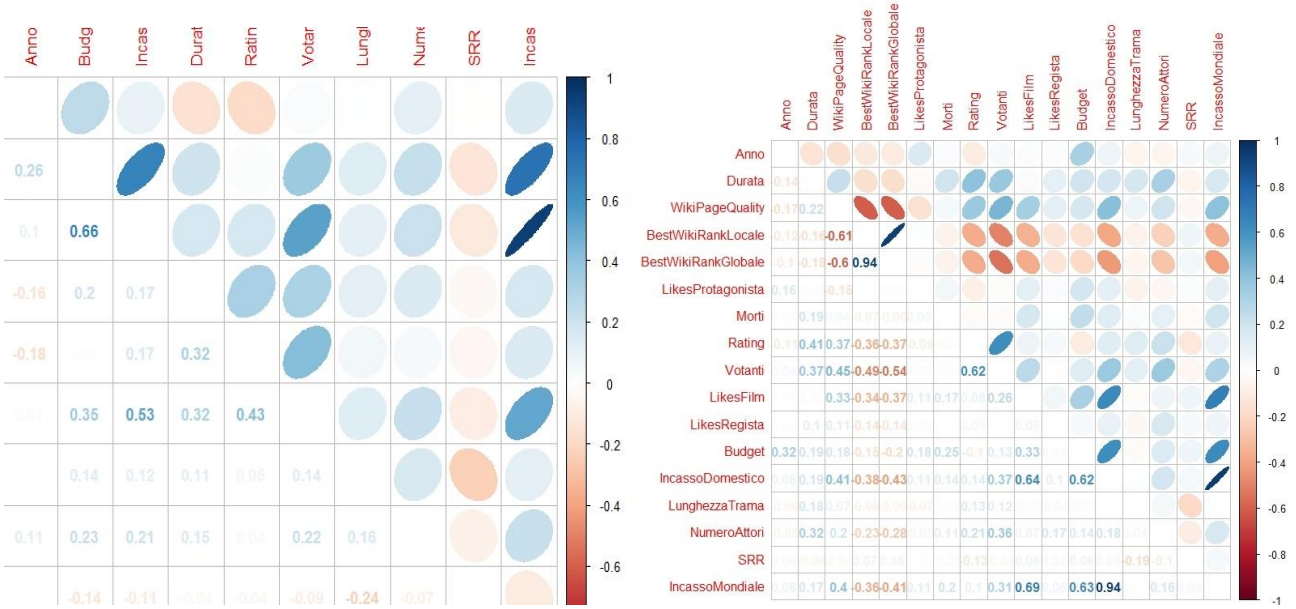
ANALISI

ANALISI DELLE CORRELAZIONI

GRAFICI DEI DATASET

Essendo di dimensioni proibitive, i grafici dei dataset sono stati allegati sottoforma di jpeg ad alta risoluzione (Immagini\ScatterPlotFilm.jpeg, Immagini\ScatterPlotFilmEsteso.jpeg).

GRAFICI DELLA CORRELAZIONE



OSSERVAZIONI

Dai grafici si può osservare come il campionamento del secondo dataset mantenga le stesse distribuzioni del primo, contenete un ordine di grandezza superiore di osservazioni, rimanendo perciò statisticamente valido.

CORRELAZIONI CON L'INCASSO MONDIALE

Le correlazioni tra l'incasso mondiale e fattori strutturali di un film, come anno d'uscita e durata, sono molto lievi e statisticamente irrilevanti (< 0.2). (Immagini\Anno-IncassoMondiale.jpeg, Immagini\Durata-IncassoMondiale.jpeg)

Le correlazioni con le statistiche della pagina *Wikipedia* sono invece non banali (± 0.4): l'incasso mondiale è positivamente correlato con la qualità della pagina di Wikipedia; negativamente correlato con il ranking massimo raggiunto a livello globale e locale. Ciò può essere dovuto al fatto che informazioni di un film di successo vengono cercate più frequentemente ed essendo maggiormente vista, più utenti sono in grado di contribuire alla qualità della pagina del film. (Si veda le immagini sottostanti 1,2,3)

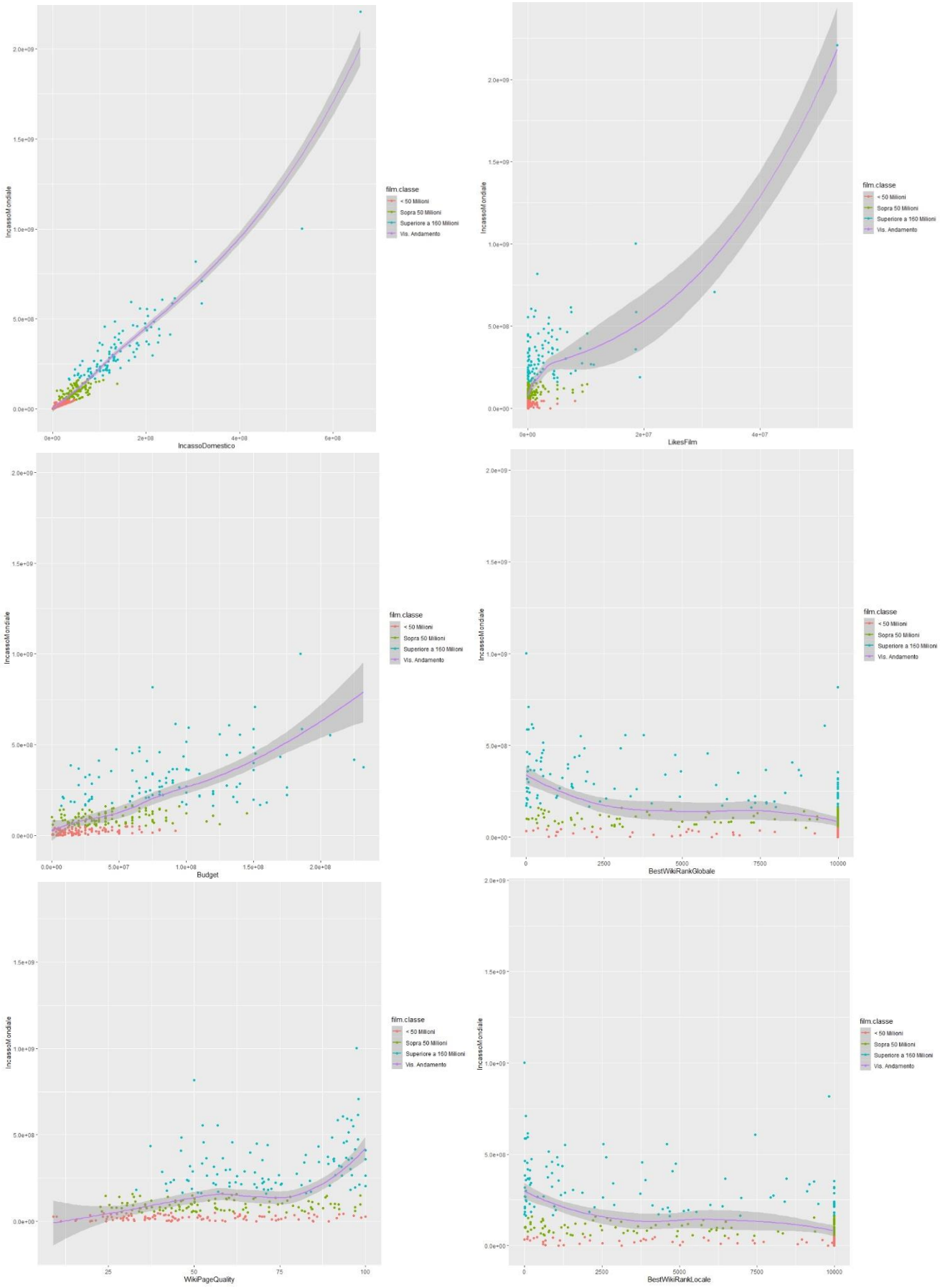
Dal punto di vista *social* l'unica correlazione netta è presente tra il numero di likes del film (± 0.7), mentre la popolarità di regista e protagonista non sono fortemente correlati (< 0.2). La popolarità del film risulta pertanto un fattore fondamentale per prevederne gli incassi globali. (Immagini\LikesProtagonista-IncassoMondiale.jpeg, Immagini\LikesRegista-IncassoMondiale.jpeg e immagine sottostante)

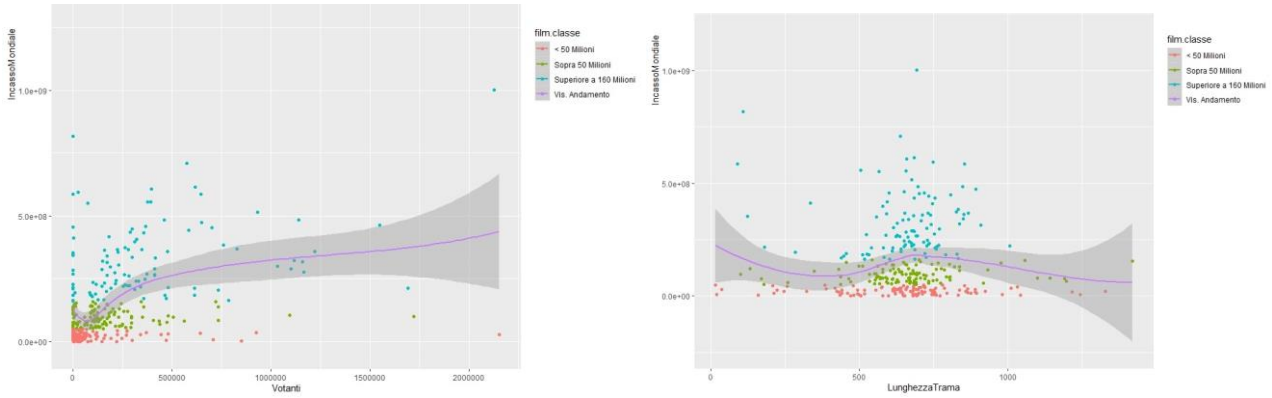
Il rating non risulta correlato con gli incassi, mentre è presente una lieve ma non trascurabile correlazione con il numero di votanti (> 0.3). Un'ipotesi è che sia più importante l'*engagement* di un film che l'apprezzamento da parte del pubblico. Può essere preso come esempio il recente e controverso film, campione di incassi, *Joker*. (Immagini\Rating-IncassoMondiale.jpeg e immagine sottostante)

La correlazione dei fattori economici con l'incasso globale risulta evidente. Un film con budget alto generalmente porta ad incassi maggiori (corr > 0.6). Gli incassi nel paese d'origine del film sono un'indicazione estremamente utile per prevedere quelli a livello globale (corr > 0.9). (Si veda le immagini sottostanti 1,2)

Fattori “pulsionali” come la percentuale di riferimenti sessuali nella trama non sembrano avere una correlazione rilevante con gli incassi del film. È presente però una correlazione che potrebbe essere non casuale con il numero di morti, anche se molto ridotta (± 0.2). (Immagini\SRR-IncassoMondiale.jpeg, Immagini\Morti-IncassoMondiale.jpeg)

Infine, è presente un’interessante correlazione a forma di imbuto rovesciato tra la lunghezza della trama e gli incassi, purtroppo non utile ai fini dell’analisi. (Si veda l’immagine sottostante)





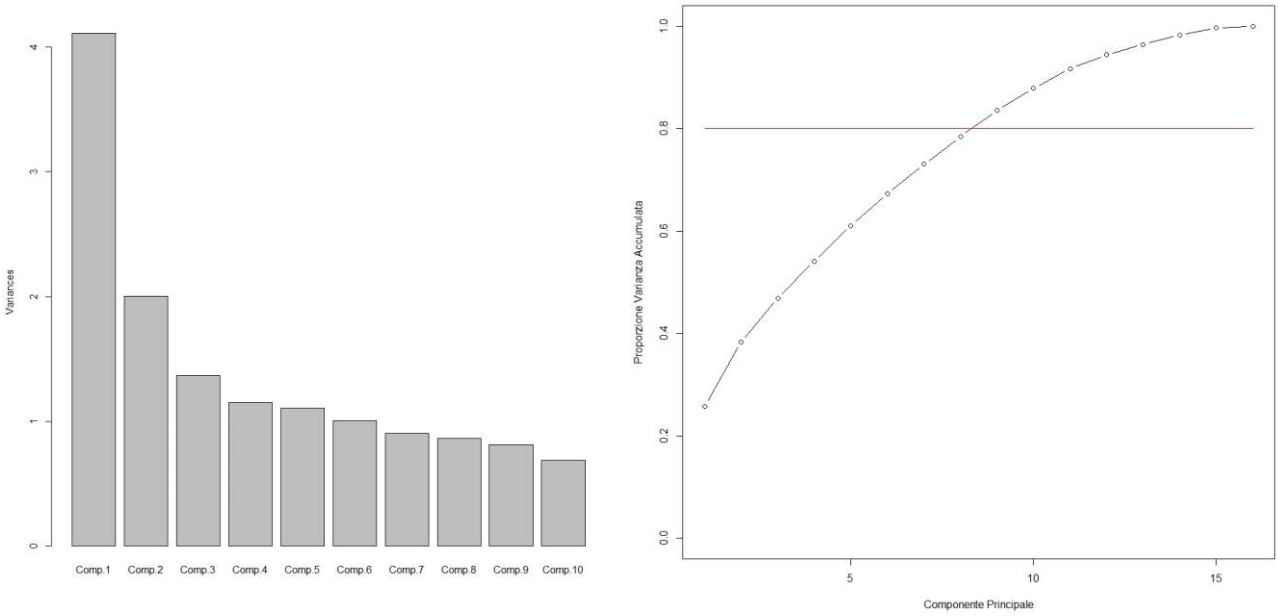
ANALISI DELLE COMPONENTI PRINCIPALI

SCOPO DELL'ANALISI

Si vuole conoscere quali fattori inducano la maggior variazione negli incassi globali di un film in modo da proiettare nelle sale quelli aventi le caratteristiche migliori, con il fine di massimizzare i guadagni.

Ai fini dell'analisi il fattore degli incassi mondiali è escluso dal calcolo delle componenti principali. I dati sono stati classificati in tre categorie con numero omogeneo di osservazioni per rappresentare gli incassi.

GRAFICI DELLA VARIANZA




```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     2.0271774 1.4146666 1.16992935 1.07368282 1.05191815 1.00220175 0.95058197 0.93049008
Proportion of Variance 0.2576308 0.1254650 0.08580914 0.07227137 0.06937103 0.06296868 0.05664915 0.05427974
Cumulative Proportion  0.2576308 0.3830958 0.46890489 0.54117626 0.61054729 0.67351597 0.73016512 0.78444486

                          Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Comp.14    Comp.15     Comp.16
Standard deviation     0.90146666 0.82908468 0.78612084 0.65356519 0.57641494 0.54030971 0.46296333 0.233713639
Proportion of Variance 0.05094641 0.04309353 0.03874297 0.02677886 0.02082978 0.01830205 0.01343716 0.003424383
Cumulative Proportion  0.83539127 0.87848480 0.91722776 0.94400662 0.96483641 0.98313846 0.99657562 1.000000000
```

Eseguendo l'analisi delle componenti principali sui fattori normalizzati, emerge che solo le prime due componenti si distinguono nettamente per varianza spiegata. Per raggiungere una varianza cumulativa dell'80% circa sono necessarie almeno otto componenti, rendendo l'analisi di difficile interpretazione.

COMPOSIZIONE DELLE COMPONENTI

Analizzando un'opportuna rotazione dei *loadings* delle prime 8 componenti principali si possono attribuire i contributi dei vari fattori rispetto alle componenti.

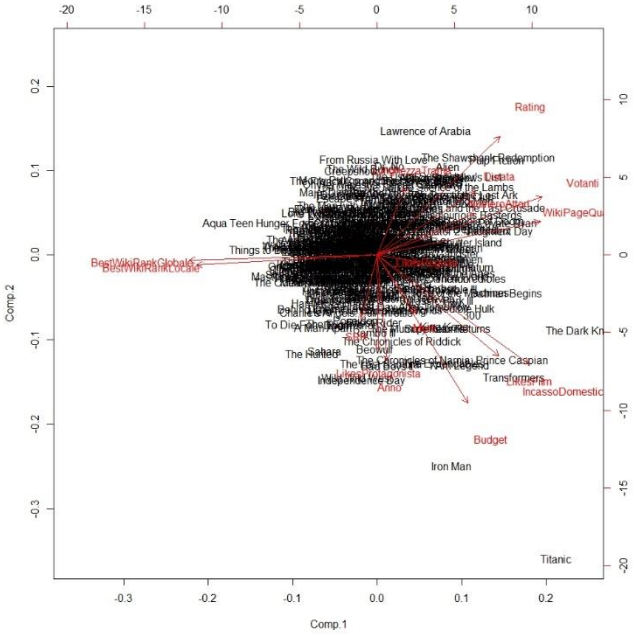
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
WikiPageQuality	0.343	-0.188		0.239			0.212	
BestWikiRankLocale	-0.591							
BestWikiRankGlobale	-0.560							
Anno	0.151			-0.834				
LikesProtagonista		-0.109					-0.898	
Durata	-0.254		-0.515		-0.202	0.218		
Morti						0.838		
Rating		0.109	-0.546	0.110				
Votanti	0.144		-0.525					
LikesFilm	0.127	-0.531	0.109	0.217			-0.104	
LikesRegista			0.177		-0.882	-0.139		
Budget	-0.117	-0.470		-0.392		0.158		
IncassoDomestico		-0.590						
LunghezzaTrama	-0.128	-0.144				-0.186	0.187	0.782
NumeroAttori			-0.278		-0.387	0.287		
SRR	-0.209	-0.209	-0.101	-0.102		-0.259	0.295	-0.613

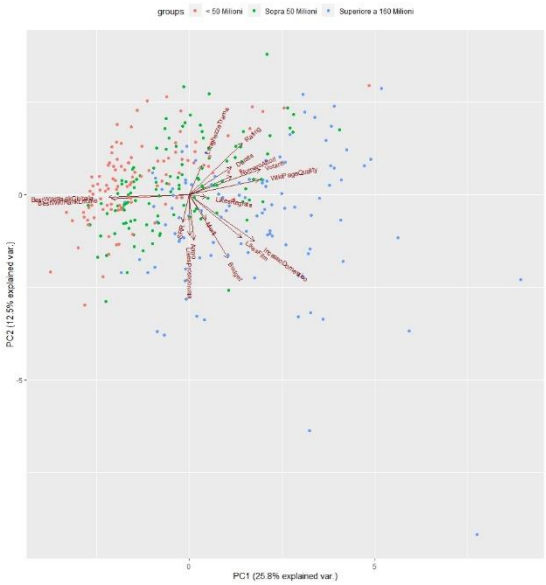
La prima componente sembra essere influenzata maggiormente da fattori collegati alla pagina di Wikipedia (*WikiPageQuality*, *BestWikiRankLocale* e *BestWikiRankGlobale*).

La seconda componente invece sembra essere governata da fattori economici (*IncassoDomestico* e *Budget*) insieme alla popolarità del film (*LikesFilm*).

La terza componente è legata al rating e al numero di votanti, ma stranamente anche alla durata della pellicola.
Le componenti successive sono influenzate dai seguenti singoli fattori in ordine: *Anno*, *LikesRegista*, *Morti* e *LikesProtagonista*, mentre l'ottava componente è legata alla trama e alla percentuale di riferimenti sessuali presenti nella trama stessa.

BIPLOT





Analizzando i *biplot* è evidente come i film di maggior successo (in azzurro) tendano a distribuirsi verso l'angolo in basso a destra, mentre i film con incassi nella norma o esigui (in verde e rosso) si distribuiscono perlopiù sul lato in alto a sinistra. Possono essere presi come esempio *Trasformers*, *Iron Man* e *The Dark Knight*, film campioni di incassi. La tendenza risulta ancora più evidente se si osserva la posizione del film, che fino a poco tempo fa era considerato il più grande successo della storia, *Titanic*, posizionato nell'estremo angolo in basso a destra.

La direzione delle frecce dei fattori relativi alla pagina Wikipedia conferma il loro contributo prevalente sulla variabilità della prima componente principale. *L'IncassoDomestico*, il *Budget* e i *LikesFilm* sembrano essere invece fattori trasversali alle componenti, la cui direzione e allineata con il verso della distribuzione di film a maggior profitto.

Il *biplot* relativo alla prima e seconda componente (Immagini\biplotFilmClasse2.jpeg) e quello relativo alla seconda e terza componente (Immagini\biplotFilmClasse3.jpeg) iniziano già a contenere osservazioni di classi differenti mescolate, risultando poco utili per la distinzione dei film di successo.

CONCLUSIONE

Le direzioni relative a *IncassoDomestico*, *Budget* e i *LikesFilm* si allineano alla distribuzione dei film di successo, unite ai fattori relativi alla pagina Wikipedia permettono di distinguere, anche se non in maniera netta, le varie classi di introiti ricavabili dalla proiezione della pellicola.

REGRESSIONE LINEARE

SCOPO DELL'ANALISI

Si vuole prevedere, con la miglior accuratezza, il possibile incasso globale di un film, in modo da preferire la proiezione delle pellicole che garantiscono il maggior profitto.

ANALISI MODELLO NON OTTIMIZZATO

La regressione lineare multivariata risulta avere una percentuale di *varianza spiegata aggiustata* superiore al 90% e un *p-value* molto basso.

L'incasso nel paese d'origine e la popolarità del film, insieme al budget di produzione, sono i fattori che catturano meglio la varianza del problema, come era emerso già nell'analisi delle componenti principali.

Contrariamente da quanto emerso precedentemente i fattori relativi alla pagina di Wikipedia sono sostituiti da quelli relativi ai likes del regista, al numero di votanti e al numero di morti nel film, nel determinare gran parte del resto della variabilità.

```
Residuals:

       Min          1Q      Median          3Q         Max
-234020877  -24737986     593955   23620608  420526433

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)

(Intercept)         -8.019e+07  7.754e+08  -0.103  0.91769
WikiPageQuality      3.046e+05  2.005e+05   1.519  0.12974
BestWikiRankLocale   4.813e+02  2.429e+03   0.198  0.84304
BestWikiRankGlobale -5.499e+02  2.531e+03  -0.217  0.82816
Anno                 1.766e+04  3.837e+05   0.046  0.96332
LikesProtagonista   -1.685e-02  2.082e-01  -0.081  0.93556
Durata               3.824e+04  1.602e+05   0.239  0.81155
Morti                8.995e+04  4.080e+04   2.205  0.02821 *
Rating               1.344e+04  4.446e+05   0.030  0.97591
Votanti             -3.743e+01  1.578e+01  -2.372  0.01829 *
LikesFilm            6.310e+00  9.970e-01   6.328 8.69e-10 ***
LikesRegista        -8.228e+00  3.927e+00  -2.096  0.03693 *
Budget               2.979e-01  1.073e-01   2.776  0.00584 **
IncassoDomestico     2.077e+00  7.300e-02  28.448  < 2e-16 ***
LunghezzaTrama       1.191e+04  1.737e+04   0.685  0.49364
NumeroAttori         6.082e+05  1.597e+06   0.381  0.70367
SRR                 -9.865e+05  2.688e+06  -0.367  0.71388

Residual standard error: 58450000 on 309 degrees of freedom
Multiple R-squared:  0.9129,    Adjusted R-squared:  0.9084
F-statistic: 202.4 on 16 and 309 DF,  p-value: < 2.2e-16
```

ANALISI MODELLO OTTIMIZZATO

Con questo modello si è cercato di ottimizzare al massimo le capacità di previsione della regressione. A tale scopo si è sostituito il logaritmo dei fattori con andamento polinomiale al loro valore precedente, in modo da linearizzarne l'andamento. Il modello di seguito è la miglior combinazione trovata al fine di massimizzare le correlazioni e di conseguenza la varianza spiegata. È stato calcolato il logaritmo di *Morti* e *LunghezzaTrama*; il logaritmo di *LikesProtagonista* + 1, *LikesRegista* + 1, *BestWikiRankLocale* + 1 e *BestWikiRankGlobale* + 1 (a causa della presenza di zeri).

```
Residuals:

       Min          1Q      Median          3Q         Max
-212734977  -21685774    -197782   22507179  395725177

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)

(Intercept)         -7.535e+08  7.932e+08  -0.950  0.34294
WikiPageQuality      5.129e+05  1.895e+05   2.707  0.00717 **
BestWikiRankLocale   2.127e+07  8.975e+06   2.370  0.01838 *
BestWikiRankGlobale -2.093e+07  1.026e+07  -2.041  0.04215 *
Anno                 3.417e+05  3.921e+05   0.871  0.38418
LikesProtagonista   -9.193e+06  4.049e+06  -2.271  0.02386 *
Durata               7.331e+04  1.601e+05   0.458  0.64738
Morti                1.894e+05  5.911e+04   3.203  0.00150 **
Rating              -2.403e+04  4.351e+05  -0.055  0.95598
Votanti             -4.090e+01  1.560e+01  -2.622  0.00917 **
LikesFilm            6.409e+00  1.003e+00   6.392 6.05e-10 ***
LikesRegista        -7.571e+00  3.847e+00  -1.968  0.04992 *
Budget               2.510e-01  1.069e-01   2.347  0.01957 *
IncassoDomestico     2.088e+00  7.195e-02  29.017  < 2e-16 ***
LunghezzaTrama       7.657e+06  7.383e+06   1.037  0.30050
NumeroAttori         4.624e+05  1.575e+06   0.294  0.76930
SRR                 -4.053e+05  2.797e+06  -0.145  0.88487

Residual standard error: 57280000 on 309 degrees of freedom
Multiple R-squared:  0.9163,    Adjusted R-squared:  0.912
F-statistic: 211.5 on 16 and 309 DF,  p-value: < 2.2e-16
```

La nuova regressione risulta avere una percentuale di *varianza spiegata aggiustata* superiore al 91% e un *p-value* sempre molto basso. Inoltre, la distribuzione dei residui è leggermente più omogenea rispetto al precedente modello.

Da notare come i fattori relativi alla pagina di Wikipedia, nel nuovo modello, concorrano nel determinare una parte non trascurabile della variabilità, come era emerso nell'analisi delle componenti principali.

ANALISI MODELLO SEMPLIFICATO

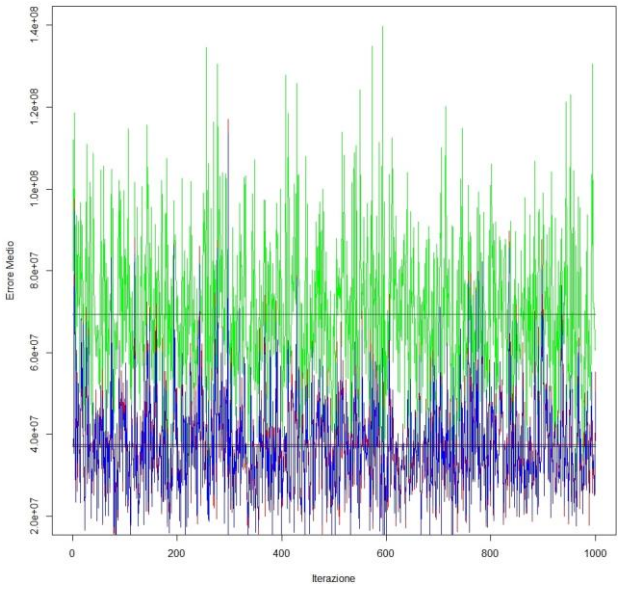
In quest'ultimo modello si è cercato di ridurre al minimo la complessità, mantenendo una percentuale di *varianza spiegata aggiustata* superiore al 90% e un equilibrio tra le importanze relative dei fattori nel determinare la variabilità del modello. A tale scopo, eliminando uno ad uno i fattori meno influenti, si è giunti al modello a fianco.

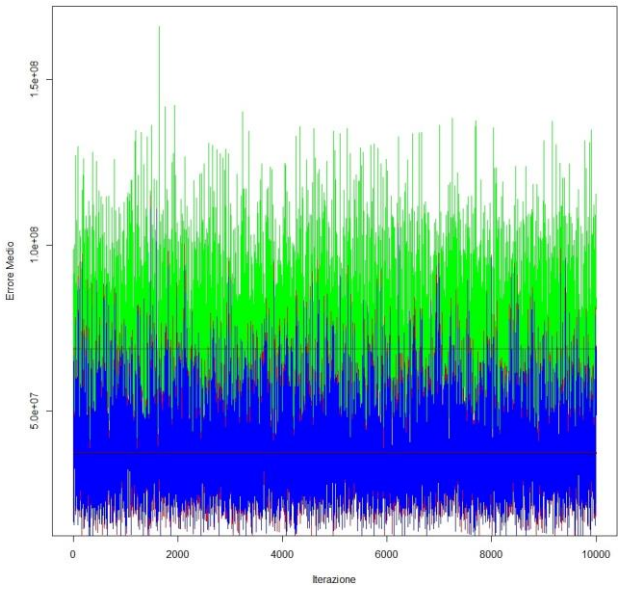
```
Residuals:

       Min         1Q     Median         3Q        Max
-253418401  -23217776     317950   18320523  462211498

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.979e+07  5.100e+06  -3.882 0.000126 ***
LikesFilm         6.607e+00  9.648e-01   6.847 3.82e-11 ***
Budget            3.724e-01  9.203e-02   4.047 6.51e-05 ***
IncassoDomestico  2.031e+00  6.820e-02  29.781  < 2e-16 ***

Residual standard error: 59030000 on 322 degrees of freedom
Multiple R-squared:  0.9074,     Adjusted R-squared:  0.9065
F-statistic:  1052 on 3 and 322 DF,  p-value: < 2.2e-16
```

Come evidenziato nell'analisi delle componenti principali *LikesFilm*, *Budget* e *IncassoDomestico* risultano essere i fattori più utili nella previsione degli incassi mondiali, avendo da soli una varianza spiegata superiore al 90%.

CONFRONTO TRA MODELLI

Sono stati svolti due esperimenti in modo da stabilire l'effettiva precisione dei modelli, ed effettuare un confronto tra di essi. I test prevedevano un training set di 316 osservazioni e un test set di 10 osservazioni. È stato calcolato l'errore medio assoluto delle previsioni sulle osservazioni, ripetendo il calcolo con campionamenti diversi rispettivamente 1000 e 10000 volte. È stata poi evidenziata la media di tutti gli errori medi assoluti per ogni modello (non ottimizzato in rosso, ottimizzato in blu e ridotto in verde).



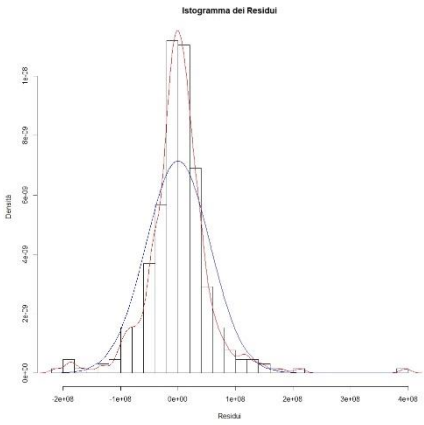


Gli errori medi risultano essere in media di 37 milioni di dollari per i primi due modelli, con una previsione migliore di mezzo milione per il modello ottimizzato. L'errore medio sale a circa 69 milioni di dollari per il modello ridotto, circa l'86% più alto rispetto agli altri modelli.

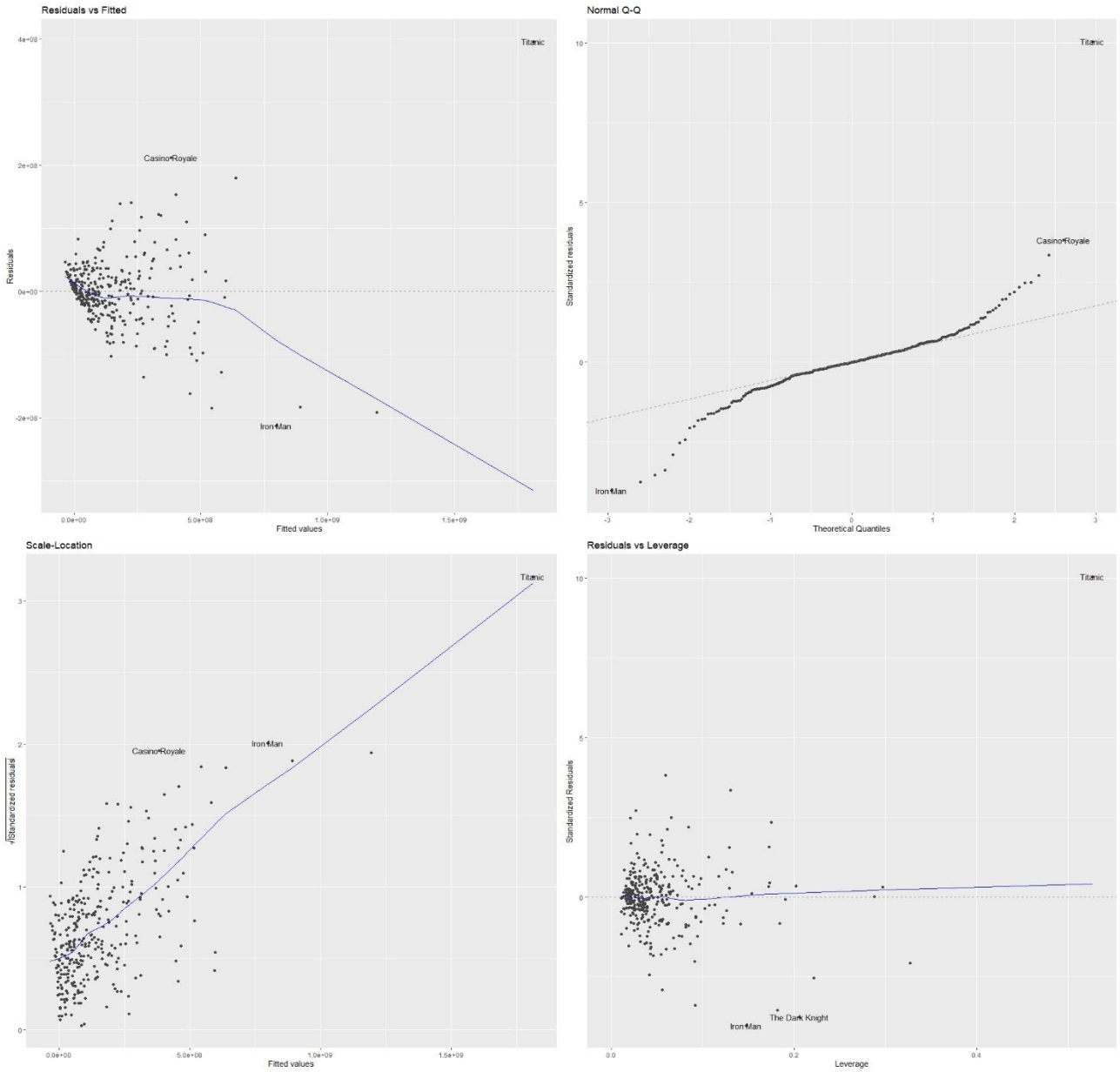
ANALISI DEI RESIDUI

Dato che lo scopo dell'analisi era ottenere la miglior previsione possibile, vengono evidenziante ulteriori osservazioni esclusivamente per il modello ottimizzato.

Analizzando l'istogramma dei residui a fianco, emerge come essi tendano ad avere una distribuzione più schiacciata rispetto ad una distribuzione normale. Infatti, eseguendo il *test di normalità Shapiro-Wilk* risulta un *p-value* molto elevato, rigettando l'ipotesi di normalità.



Si rendono perciò necessarie ulteriori analisi sulla distribuzione dei residui:

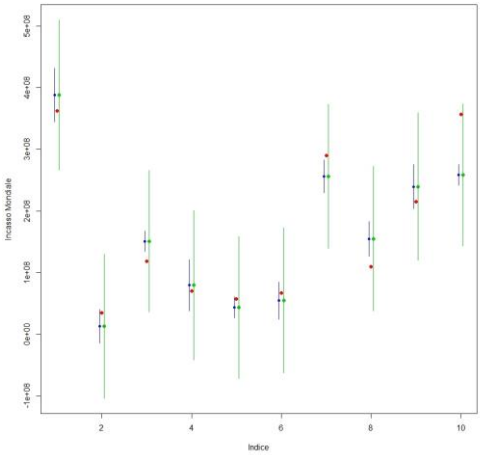


Dai grafici risulta evidente che la qualità della predizione diminuisce all'aumentare degli incassi effettivi di un film, ciò può essere dovuto alla presenza di svariati *outlier* tra cui spicca *Titanic*, caratterizzato da un *leverage* molto elevato. Ciò spiega anche l'andamento schiacciato della distribuzione dei residui nell'istogramma.

Non sembrano evidenti altri tipi di influenze sulla previsione, come per esempio dovute a relazioni di tipo non lineare tra i fattori.

CONCLUSIONE

Le previsioni del modello, come evidenziato nella figura a fianco, rientrano generalmente all'interno dei *margini di previsione* per valori di incassi non elevati e all'interno dei *margini di confidenza* per gli incassi più elevati. In ogni caso non sembrano verificarsi scambi nell'ordinamento globale delle previsioni rispetto all'ordine dei valori reali. La minor precisione sui valori più elevati sembra essere naturalmente spiegabile dal fatto che film campioni di incassi tendono ad essere un'eccezione (*outlier*) rispetto alla norma e quindi più difficili da prevedere.



Il modello risulta perciò adatto alla previsione e ordinamento dei film in base al loro incasso globale.