Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

Relazione tra l'esito del test di ingresso TOLC_I e le carriere universitarie: una analisi statistica basata su alcuni corsi della Scuola di Scienze

Relatrice Prof.ssa Laura Ventura
Dipartimento di Scienze Statistiche

Correlatrice Prof.ssa Camilla Ferrante
Dipartimento di Scienze Chimiche

Correlatrice Dott.ssa Marta Molena
Scuola di Scienze

Laureando Tiziano Cicerchia
Matricola N. 1233218

ANNO ACCADEMICO 2022/23

per chi vorrebbe ma non può
28/03/1960

Indice

Introduzione

Le analisi contenute in questo elaborato derivano da una esperienza di stage svolta presso la Scuola di Scienze dell'Università degli Studi di Padova. Il lavoro è stato supervisionato dalla Dott.ssa Marta Molena (Scuola di Scienze) e dalla Prof.ssa Camilla Ferrante (Presidente della Commissione Test d'Ingresso della Scuola di Scienze).
Le analisi sono volte a valutare l'eventuale presenza di relazioni tra i punteggi ottenuti al TOLC_I e le carriere universitarie di un campione di studenti immatricolati nel 2019 ad alcuni corsi di laurea della Scuola di Scienze dell'Università degli Studi di Padova. Oltre alla relazione marginale tra il punteggio generale ottenuto al test e il voto di laurea, vengono studiate le relazioni che intercorrono tra i punteggi ottenuti nelle singole sezioni che compongono il TOLC_I e altri aspetti della carriera universitaria degli studenti, come la durata del percorso di studi, il numero di esami sostenuti e le medie dei voti.
TOLC è l'acronimo di Test OnLine CISIA (https://www.cisiaonline.it/area-tematica-tolc-cisia/cose-il-tolc/). È un test per chi vuole iscriversi a un corso di laurea che richiede una valutazione delle conoscenze iniziali prima dell'iscrizione. Le conoscenze oggetto di valutazione dipendono dal corso di laurea scelto. Il TOLC è diverso da studente a studente ed è composto da quesiti selezionati dal database CISIA TOLC. Tutti i TOLC appartenenti alla medesima tipologia hanno un livello di difficoltà analogo o comunque paragonabile. Può essere usato anche come test di selezione per i corsi di laurea ad accesso programmato locale. In base al risultato conseguito nel TOLC, le università possono indicare agli studenti e alle studentesse quali corsi integrativi seguire e attribuire degli OFA (Obblighi Formativi Aggiuntivi) da colmare oppure stabilire delle propedeuticità all'interno degli esami curriculari [1]. Esistono diversi tipi di TOLC strutturati in maniera diversa, sia per numero sia per tipo e difficoltà di quesiti.
Nelle analisi che seguono sono stati presi in considerazione studenti che hanno sostenuto il TOLC_I nel 2019.
Questa tipologia di test è costituita da 50 quesiti suddivisi in 4 sezioni: Matematica, Logica, Scienze, Comprensione Verbale. Il risultato di ogni TOLC_I, ad esclusione della sezione relativa alla prova della conoscenza della Lingua Inglese, è determinato dal numero di risposte esatte, sbagliate e non date che determinano un punteggio assoluto, derivante da 1 punto per ogni risposta corretta, 0 punti per ogni risposta non data e una penalizzazione di 0,25 punti per ogni risposta errata.
Ai fini del calcolo del punteggio per l'immatricolazione ai Corsi di Laurea della Scuola di Scienze di Padova e per l'attribuzione degli OFA i punteggi sono però pesati, in base ai seguenti criteri:

- vengono considerate solo le sezioni di linguaggio matematico di base, scienze, logica e comprensione del testo;
- la sezione di Lingua inglese non viene considerata;
- per la sezione di scienze si applica peso pari a 0.1 ad ogni risposta, mentre per tutti gli altri quesiti il peso assegnato a ogni risposta è pari a 1;

- il punteggio attribuito risulta quindi al massimo 41.

Gli studenti che conseguono un punteggio inferiore a 17/41 si immatricolano, ma conseguono un OFA in Matematica. Per informazioni sul recupero degli obblighi formativi aggiuntivi si rinvia al sito della Scuola di Scienze (https://www.scienze.unipd.it/come-iscriversi/come-iscriversi-l-1920/ofa/).

Schema della tesi

L'elaborato è composto da una prima parte in cui viene descritto il dataset su cui si basano le analisi: la sua struttura, le variabili e le operazioni compiute su di esse.
In seguito vengono condotte le analisi esplorative, attraverso le quali vengono illustrate le distribuzioni empiriche delle variabili e ne vengono messe in risalto le caratteristiche salienti. In un primo momento vengono riportate le distribuzioni marginali delle singole variabili, successivamente l'attenzione si sposta sulle analisi bivariate, che permettono di osservare le relazioni che intercorrono tra coppie di variabili. Viene posta particolare attenzione al comportamento delle singole variabili condizionate al corso di laurea e al sesso.
Infine, vengono adattati modelli di regressione per studiare l'effetto congiunto di più variabili su una risposta. In questo elaborato vengono prese in considerazione come risposte il voto di laurea, il numero di esami sostenuti e la variabile che indica se uno studente è laureato o meno.
Nell'ultima sezione dell'elaborato vengono commentati e interpretati i risultati ottenuti.

Capitolo 1

Descrizione dei dati

1.1 Selezione del campione e pulizia dei dati

I dati a cui fanno riferimento le analisi sono stati forniti dalla Scuola di Scienze dell'Università degli Studi di Padova. Il dataset utilizzato per svolgere le analisi è stato ottenuto tramite il merging di tre dataset contenenti rispettivamente informazioni sul TOLC_I, voti degli esami sostenuti durante il percorso universitario e informazioni sulle lauree di studenti immatricolati nel 2019 ad alcuni corsi della Scuola di Scienze. I dati sono aggiornati alla sessione di laurea di Marzo 2023.
Dal dataset contenente le informazioni sul percorso scolastico superiore degli studenti e i punteggi conseguiti al TOLC_I sono stati selezionati i soli studenti che si sono immatricolati ai corsi di Fisica, Chimica Industriale, Matematica, Scienze Naturali e Scienze e Tecnologie per l'Ambiente (STAM). I corsi selezionati sono stati concordati con la Commissione Test d'Ingresso della Scuola di Scienze.
Il secondo dataset, contenente i voti degli esami sostenuti da ciascuno studente, è stato riorganizzato mediante l'utilizzo del software `SAS`. È stato necessario riorganizzare il dataset perchè i dati erano inizialmente disposti in modo da avere una riga per ogni esame sostenuto dal singolo studente. In seguito all'operazione svolta in `SAS`, ogni riga del dataset fa riferimento ad uno studente e per ciascuno di essi sono registrati i voti ed il numero degli esami sostenuti. Tale organizzazione dei dati è stata necessaria per rendere possibile il merging dei tre datasets. La descrizione dei datasets ed il codice utilizzato per il merging di essi è riportata in Appendice.

1.2 Descrizione delle variabili

Il dataset finale è composto da 458 unità statistiche (studenti), per ciascuna delle quali sono state rilevate le 18 variabili che seguono:

- `CF`, indica il codice fiscale degli studenti. Non è di interesse ai fini delle analisi, ma rappresenta un identificatore univoco per ogni studente.
- `sesso`, variabile qualitativa con modalità 'Maschio' e 'Femmina'.
- `superiore`, variabile qualitativa che indica il tipo di scuola superiore frequentata dal singolo studente. Inizialmente questa variabile ha 17 modalità.
- `provincia`, variabile qualitativa che indica la provincia in cui il singolo studente ha frequentato le scuole superiori.

- `regione`, variabile qualitativa che indica la regione in cui il singolo studente ha frequentato le scuole superiori. Inizialmente questa variabile ha 22 modalità.
- `maturita`, variabile quantitativa discreta che rappresenta il voto conseguito all'esame di maturità.
- `tolc`, variabile quantitativa che indica il punteggio conseguito al TOLC_I. Come descritto nel capitolo introduttivo, il punteggio massimo è pari a 41 e gli studenti che conseguono un punteggio minore di 17/41 conseguono l'OFA in Matematica.
- `comprensione`, variabile quantitativa che indica il punteggio conseguito nella sezione di comprensione del testo del TOLC_I.
- `matematica`, variabile quantitativa che indica il punteggio conseguito nella sezione di matematica del TOLC_I.
- `logica`, variabile quantitativa che indica il punteggio conseguito nella sezione di logica del TOLC_I.
- `scienze`, variabile quantitativa che indica il punteggio conseguito nella sezione di scienze del TOLC_I.
- `corso`, variabile qualitativa che indica il corso a cui lo studente si è immatricolato.
- `numeroesami`, variabile quantitativa discreta che indica il numero di esami sostenuti dal singolo studente.
- `aritmetica`, variabile quantitativa che rappresenta la media aritmetica degli esami sostenuti, disponibile per ogni studente.
- `ponderata`, variabile quantitativa che rappresenta la media ponderata degli esami sostenuti dal singolo studente. Il valore di questa variabile è disponibile solo per gli studenti che si sono laureati, per gli altri studenti sono stati inseriti valori mancanti.
- `votolaurea`, variabile quantitativa che indica il voto di laurea conseguito dagli studenti che hanno completato il percorso di studi.
- `lode`, variabile dicotomica che indica se il singolo laureato ha conseguito la lode o meno.
- `laureato`, variabile dicotomica che indica se il sigolo studente si è laureato o meno.

Nel capitolo che segue vengono descritte le distribuzioni empiriche delle variabili del dataset mediante grafici, tabelle e indici di sintesi che ne evidenziano le caratteristiche principali.

Capitolo 2

Analisi esplorative

In questo capitolo sono riportate le analisi esplorative, svolte mediante l'utilizzo del software statistico R (https://www.r-project.org/). In primo luogo vengono svolte analisi univariate al fine di fornire le principali misure di sintesi della distribuzione di ciascuna variabile. Successivamente, viene fornita una più approfondita descrizione del campione mediante analisi bivariate, le quali permettono di indagare la presenza di eventuali relazioni tra variabili. Infine, vengono analizzate le distribuzioni delle variabili inerenti al TOLC_I e alla carriera universitaria stratificando per corso di laurea e per sesso. Questo tipo di analisi permette la valutazione di differenze e analogie in termini di rendimento tra i 5 corsi e in base al sesso degli studenti. A causa dell'elevato numero di variabili, in questo capitolo vengono riportate solo le analisi significative per non appesantire l'elaborato.

2.1 Analisi univariate

Il campione è composto prevalentemente da studenti di sesso maschile, la cui frequenza percentuale è pari al 61.57%, mentre per l'intera Scuola di Scienze tale frequenza è del 55%. In Figura 2.1 sono riportate le frequenze assolute della variabile `sesso`.
Per quanto riguarda il corso di studi, si nota che il 36.0% degli studenti sono immatricolati al corso di Fisica e il 28.8% a Matematica, mentre i rimanenti sono suddivisi in proporzioni simili tra gli altri tre corsi. Le frequenze percentuali di iscritti a ciascun corso di laurea sono riportate in Figura 2.2.

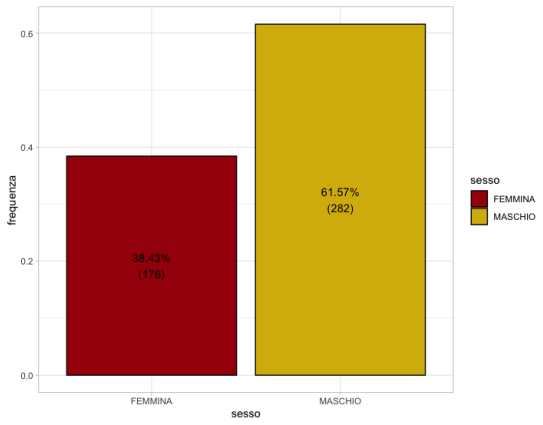

Figura 2.1: Barplot della variabile `sesso`

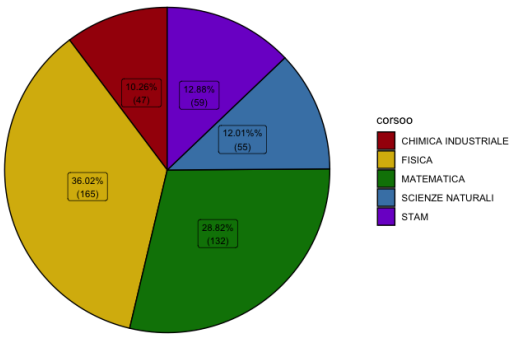

Figura 2.2: Grafico a torta della variabile `corso`

Il 72.7% (333) degli studenti ha frequentato le scuole superiori in Veneto. In corrispondenza di tutte le altre modalità della variabile `regione` si registrano frequenze percentuali minori del 6%, in alcune regioni non viene raggiunto neanche l'1%. Neanche nelle regioni limitrofe al Veneto vengono osservate frequenze di particolare interesse; nello specifico il 5.9% degli studenti (27) ha frequentato le scuole superiori in Friuli Venezia-Giulia, il 5.2% (24) in Lombardia, il 3.5% (16) in Trentino Alto-Adige e solo lo 0.9% (4) in Emilia Romagna. Alla luce di queste analisi, la variabile `regione` è stata ridefinita come una variabile dicotomica che indica se il singolo studente ha frequentato le scuole superiori in Veneto o meno. In Figura 2.3 viene raffigurata la distribuzione della variabile `regione` in seguito alla dicotomizzazione.

Per quanto riguarda la tipologia di scuola superiore frequentata (`superiore`), la situazione è molto simile a quella osservata per la variabile `regione`: il 68.1% degli studenti (312) ha frequentato il Liceo Scientifico, la seconda frequenza percentuale più alta è 12.4% in corrispondenza dell'Istituto Tecnico Industriale. Gli studenti provenienti da queste due tipologie di scuole superiori compongono circa l'80% del campione, il rimanente 20% è suddiviso tra altre 15 tipologie, in corrispondenza di ciascuna delle quali il numero di studenti è inferiore o uguale a 25. Questi risultati suggeriscono di trattare `superiore` analogamente a quanto fatto per `regione`, cioè di codificarla come una variabile dicotomica che indica se il generico studente ha frequentato il Liceo Scientifico o meno.

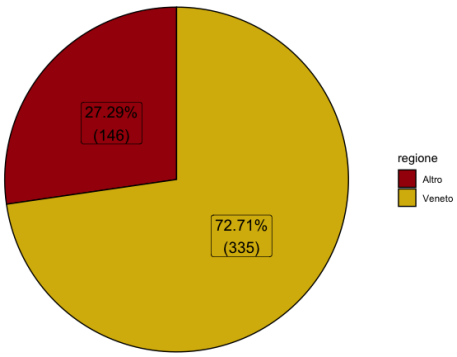

Figura 2.3: Grafico a torta della variabile `regione` dicotomizzata

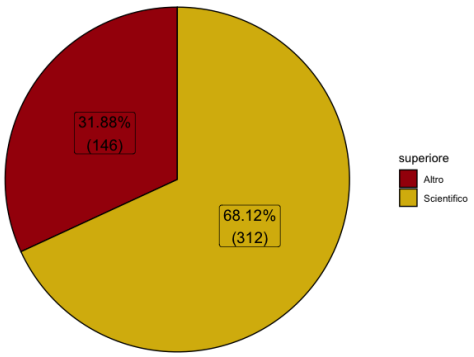

Figura 2.4: Grafico a torta della variabile `superiore` dicotomizzata

Il voto medio conseguito all'esame di maturità è circa pari a 89 (s.d.=12.26). Il boxplot in Figura 2.5 fornisce un'indicazione grafica dell'elevata variabilità del voto di maturità. La distribuzione di frequenza della variabile `maturita`, come si può osservare in Figura 2.6, è unimodale: oltre il 30% degli studenti ha superato l'esame con punteggio pari a 100/100. Tale punteggio rappresenta la moda della distribuzione. Inoltre, il 75% degli studenti ha conseguito il diploma con un voto superiore a 80/100.

Misure di sintesi sui punteggi del TOLC_I e su aspetti delle carriere universitarie

Il voto medio conseguito nel test è pari a 29.19 (s.d.=6.85). Il punteggio appare piuttosto variabile, come si nota osservando la Figura 2.7, in cui viene riportato il boxplot della variabile `tolc`. La percentuale di studenti che hanno conseguito l'OFA in matematica, ovvero coloro che hanno ottenuto un punteggio inferiore a 17/41 nel TOLC_I, è pari al 5% (24 studenti).

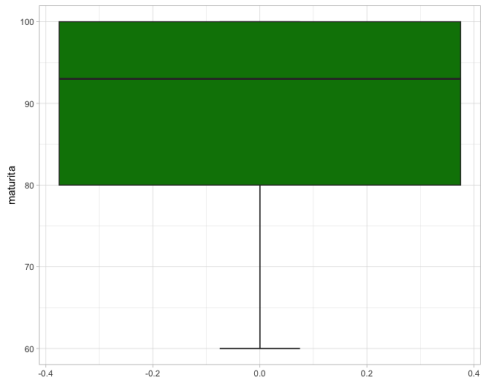

Figura 2.5: Boxplot della variabile `maturita`

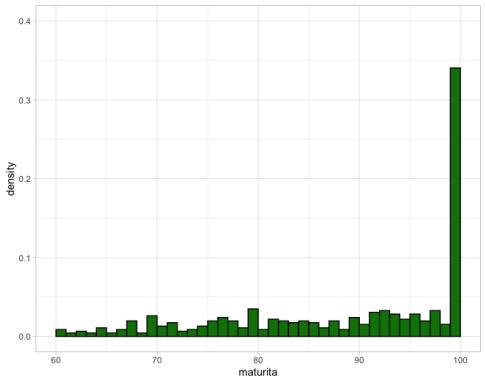

Figura 2.6: Istogramma della variabile `maturita`

I risultati nelle singole sezioni del test, analogamente al punteggio generale, appaiono piuttosto variabili. I punteggi medi ottenuti in ciascun campo e i relativi standard error sono riportati in Tabella 2.1, mentre in Figura 2.8 si possono osservare boxplot delle distribuzioni delle variabili `comprensione`, `matematica`, `logica` e `scienze`. Osservando i grafici in figura 2.8 si nota che la distribuzione dei punteggi della sezione di Comprensione del testo appare spostata verso valori più alti rispetto alle altre. Inoltre, tale distribuzione è caratterizzata da una variabilità minore rispetto alle altre anche se sono presenti tre outliers in corrispondenza della coda sinistra.
Tuttavia, il confronto tra le quattro sezioni va effettuato con cautela in quanto i punteggi sono espressi su scale diverse. Nella sezione delle analisi bivariate verranno effettuati confronti tra i punteggi all'interno delle singole sezioni stratificando per sesso e per corso di laurea.

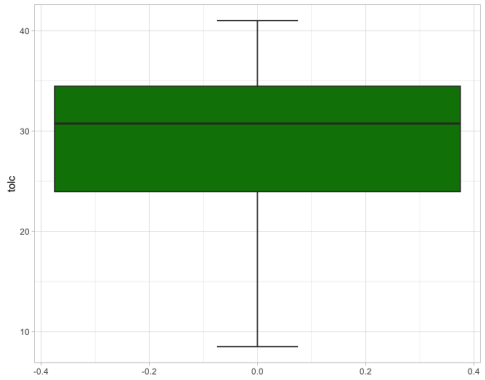

Figura 2.7: Boxplot dei punteggi ottenuti al TOLC_I

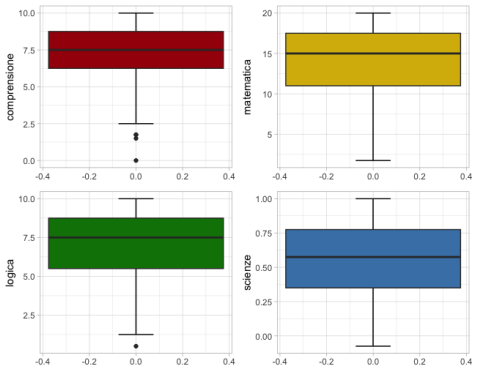

Figura 2.8: Boxplot dei putteggi ottenuti nelle sezioni del TOLC_I

	Comprensione	Matematica	Logica	Scienze	TOLC_I
Media	7.44	14.12	7.07	0.55	29.19
s.d.	1.96	4.29	2.18	0.27	6.85

Tabella 2.1: Medie e s.d. delle variabili `comprensione`, `matematica`, `logica`, `scienze` e `tolc`

Il numero medio di esami sostenuti dagli studenti è circa 14 (s.d.=6.98). Di seguito vengono riportate le caratteristiche della distribuzione di `numeroesami`. Le principali statistiche di sintesi sono riportate in Tabella 2.2. Per una rappresentazione grafica della distribuzione del numero di esami si vedano le Figure 2.9 e 2.10. Il valore mediano, pari a 19, sta a indicare che metà degli studenti ha sostenuto 19, 20 o 21 esami[1]. In particolare, il 49.24% degli studenti ha sostenuto 19 o 20 esami e sono quindi laureati o prossimi alla laurea. Osservando la Figura 2.10 si notano frequenze particolarmente elevate in corrispondenza di tali valori della variabile `numeroesami`.

Min	1st Qu.	Median	Mean	3rd Qu.	Max	s.d.
1.00	8.00	19.00	14.05	20.00	20.00	6.98

Tabella 2.2: Principali statistiche di sintesi della variabile `numeroesami`

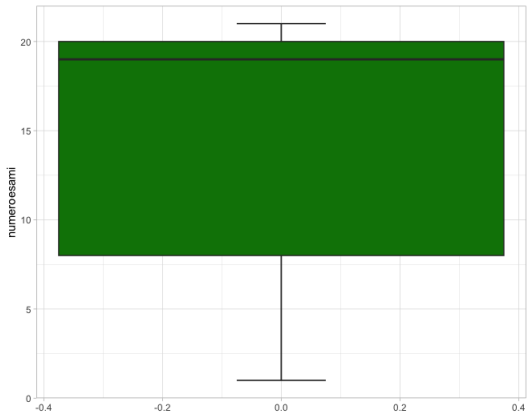

Figura 2.9: Boxplot della variabile `numeroesami`

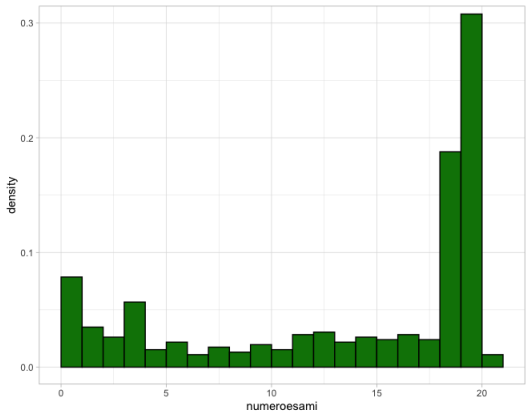

Figura 2.10: Istogramma della variabile `numeroesami`

Il valore della variabile `aritmetica` è disponibile per tutti gli studenti, al contrario della media ponderata (variabile `ponderata`) che è disponibile solo per gli studenti che si sono laureati. Il valore medio è pari 25.30 (s.d.=2.47). Nella Figura 2.11 viene riportato l'istogramma della variabile `aritmetica`.
Gli studenti che si sono laureati entro tre anni compongono il 42.36% del campione (194 studenti). Come anticipato, per gli studenti laureati è disponibile la media ponderata, la quale consiste in una media dei voti ottenuti nei singoli esami, con ciascun voto pesato per il numero di crediti assegnati all'esame. Nella Figura 2.12 si ha una rappresentazione grafica della sua distribuzione.
Per quanto riguarda la variabile `votolaurea` il valore minimo è 88/110, il massimo è 110/110. Il voto massimo rappresenta la moda della distribuzione: il 28.87% degli studenti che hanno completato il percorso universitario (56 su 196) si è laureato con 110/110. Il 75% degli studenti che si sono laureati con 110/110 ha conseguito la lode. Gli studenti che si sono laureati con lode costituiscono il 21.65% del campione.

Nella prossima sezione vengono presentate le analisi bivariate. L'obiettivo principale di queste analisi è lo studio della relazione tra i punteggi ottenuti al TOLC_I e il rendimento universitario. Viene posta particolare attenzione alla valutazione delle differenze in termini di rendimento tra studenti immatricolati a corsi di laurea diversi.

[1] Alcuni studenti hanno sostenuto esami sovrannumerari rispetto ai 20 previsti dai decreti ministeriali.

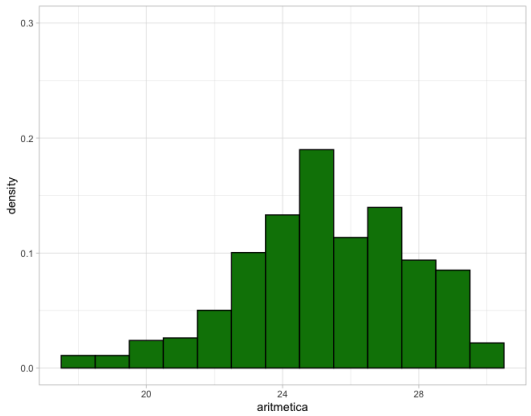

Figura 2.11: Istogramma della variabile `aritmetica`

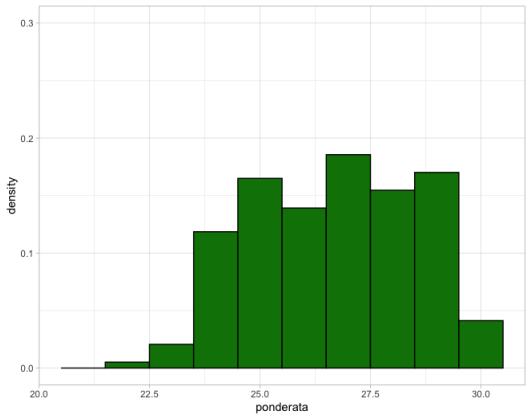

Figura 2.12: Istogramma della variabile `ponderata`

2.2 Analisi bivariate

Nella prima parte di questa sezione vengono forniti alcuni risultati utili a descrivere in modo più approfondito il campione. Per approfondimenti sulle tecniche e i test statistici utilizzati si rinvia a Ventura e Racugno [2].

In Figura 2.13 sono riportate le distribuzioni della variabile `sesso` nei diversi corsi di laurea. La proporzione di studentesse appare diversa nei 5 gruppi: nel corso di Fisica la proporzione di maschi è nettamente maggiore, nei corsi di STAM e Scienze Naturali la situazione sembra più equilibrata. Il test per il confronto delle proporzioni di studentesse nei 5 gruppi definiti dalla variabile `corso` si basa sulla statistica χ^2 di Pearson. Con una statistica test χ^2 pari a 23.99 (p-value <0.01) il test fornisce una forte evidenza contro l'ipotesi di uguaglianza delle proporzioni di studentesse nei 5 corsi, confermando quanto dedotto osservando la Figura 2.13.
Il test χ^2 di Pearson condotto all'interno dei gruppi definiti dalla variabile `corso` sottopone a verifica l'ipotesi che la proporzione di studentesse sia uguale a 0.5, ossia che la proporzione di femmine sia uguale a quella di maschi. Tale test, ripetuto per ogni corso di laurea, porta a concludere che la proporzione di studentesse è significativamente diversa da quella di studenti solo per quanto riguarda il corso di Fisica (statistica test pari a 39.76, p-value <0.01). Per quanto riguarda i corsi di laurea in Scienze Naturali e STAM, vi è una netta evidenza a favore dell'ipotesi nulla: in entrambi i gruppi la proporzione di studentesse non risulta significativamente diversa da quella di studenti. Anche nel corso di laurea in Chimica Industriale viene accettata l'ipotesi di uguaglianza delle proporzioni, anche se l'evidenza a favore dell'ipotesi nulla è più lieve rispetto ai corsi di STAM e Scienze Naturali (χ^2=3.60, p-value=0.06).
Le statistiche test e i p-value dei test χ^2 per ogni corso di laurea sono riassunti in Tabella 2.3.

	CHIMICA INDUSTRIALE	FISICA	MATEMATICA	SCIENZE NATURALI	STAM
χ^2	3.60	39.76	3.03	0.00	0.61
p-value	0.06	<0.01	0.08	1	0.43

Tabella 2.3: Statistiche test e p-value dei test χ^2 per l'uguaglianza delle proporzioni all'interno dei gruppi di `corso`

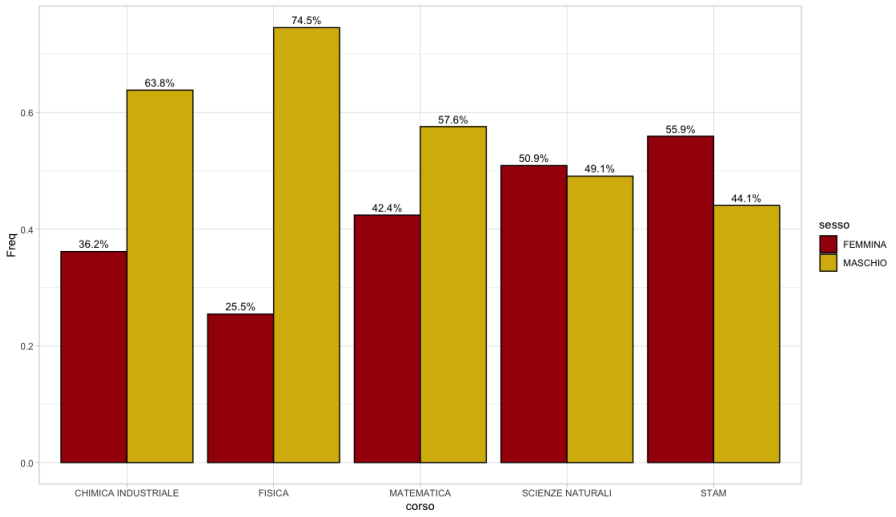

Figura 2.13: Barplot della variabile `sesso`, per `corso`, frequenze percentuali

Successivamente vengono messi a confronto i rendimenti di studenti e studentesse. In un primo momento vengono confrontati i punteggi ottenuti al TOLC_I, poi alcuni aspetti della carriera universitaria, come la media aritmetica dei voti degli esami sostenuti. In Figura 2.14 vengono messe a confronto le distribuzioni dei punteggi ottenuti al TOLC_I da maschi e femmine. La distribuzione dei punteggi conseguiti dai maschi appare spostata verso valori più alti rispetto a quella del gruppo delle femmine. Le principali statistiche di sintesi della variabile `tolc` nei gruppi definiti da `sesso` sono riportati in Tabella 2.4. L'ipotesi di normalità viene rifiutata all'interno di entrambi i gruppi: la statistica test è pari a 0.954 per i maschi e 0.975 per le femmine, il p-value è <0.01 in entrambi i casi. In Figura 2.15 sono riportati i diagrammi quantile contro quantile della variabile `tolc` nei due gruppi.
A causa della non normalità delle distribuzioni si procede al confronto dei valori mediani mediante il test non parametrico di Mann-Whitney, basato sui ranghi. In questo caso, viene testata l'ipotesi nulla che la mediana del gruppo dei maschi sia maggiore di quella del gruppo delle femmine attraverso un test unilaterale: con una statistica test pari a 30644 e p-value pari a 1 l'evidenza a favore dell'ipotesi nulla è molto forte. Si conclude quindi che il punteggio mediano del gruppo dei maschi è significativamente maggiore di quello delle femmine.

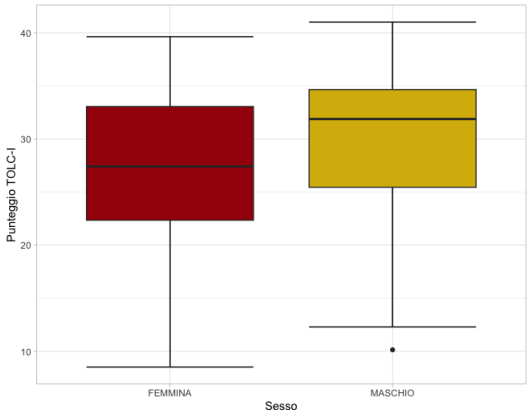

Figura 2.14: Boxplot della variabile `tolc` per i due gruppi di `sesso`

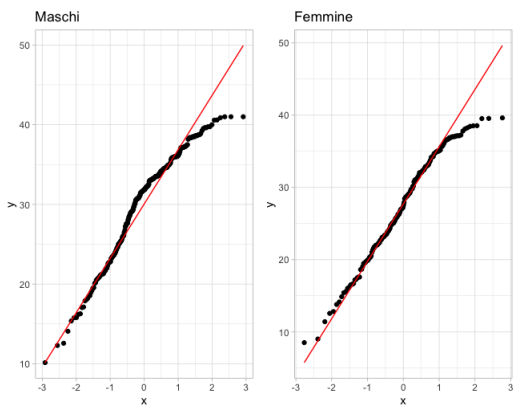

Figura 2.15: Grafico-*qq* della variabile `tolc` per i due gruppi di `sesso`

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	S.d.	IQR
Maschi	10.15	25.45	31.88	30.30	34.65	41.00	6.43	9.20
Femmine	8.53	22.35	27.413	27.409	33.05	39.63	7.11	10.70

Tabella 2.4: Principali statistiche di sintesi della variabile `tolc` per i due gruppi di `sesso`

Si considera il confronto tra i due sessi per ognuna delle quattro sezioni del test TOLC_I. L'ipotesi di normalità viene rifiutata sia per i maschi sia per le femmine per quanto riguarda tutte le sezioni del test TOLC_I, si procede quindi al confronto dei valori mediani con il test di Mann-Whitney. Alla luce dei risultati dei test, l'ipotesi di uguaglianza delle distribuzioni viene accettata per quanto riguarda la sezione di comprensione del testo. Nelle sezioni di Matematica e Scienze il punteggio conseguito dagli studenti risulta significativamente più alto rispetto a quello conseguito dalle studentesse, mentre nella sezione di Comprensione è difficile stabilire se ci sia una differenza significativa tra le distribuzioni dal momento che il p-value è pari a 0.03. In Figura 2.16 si possono osservare le distribuzioni dei punteggi conseguiti da maschi e femmine in ogni sezione del test. I p-value dei test di normalità sono riportati in Tabella 2.5, quelli dei test per il confronto dei valori mediani in Tabella 2.6. L'ipotesi nulla sottoposta a verifica con il test di Mann-Whitney è che il valore mediano della distribuzione dei maschi sia uguale a quello delle femmine.

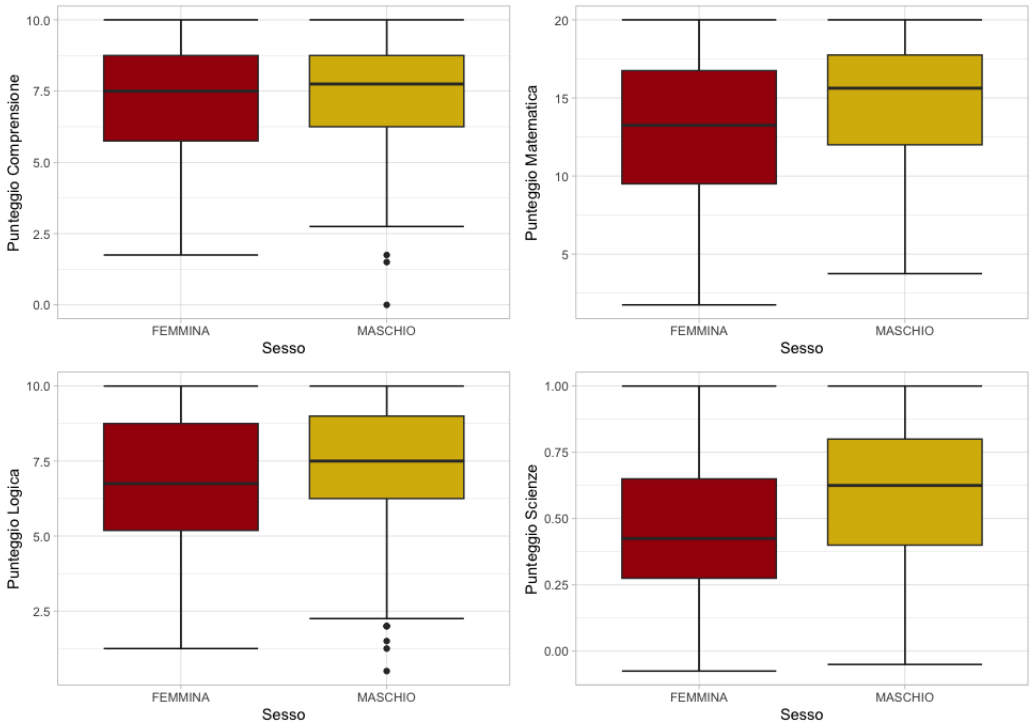

Figura 2.16: Boxplot della variabile **a)** `comprensione`, per `sesso`, **b)** `matematica`, per `sesso` **c)** `logica`, per `sesso` **d)** `scienze`, per `sesso`

Come anticipato nei capitoli precedenti, gli studenti che non raggiungono il punteggio di 17/41 nel test conseguono l'OFA in Matematica. L'8.52% delle studentesse (15 su 176) ha conseguito l'OFA, tale percentuale scende al 3.2% (9 su 282) per quanto riguarda gli studenti. Con una statistica test pari a 5.18

Test di Shapiro-Wilk				
	Maschi		Femmine	
Sezione	W_{SW}^{oss}	p-value	W_{SW}^{oss}	p-value
Comprensione	0.927	<0.01	0.928	<0.01
Matematica	0.934	<0.01	0.960	<0.01
Logica	0.936	<0.01	0.959	<0.01
Scienze	0.961	<0.01	0.977	<0.01

Tabella 2.5: Statistiche test e p-value dei test di normalità per i punteggi di maschi e femmine nelle sezioni del TOLC_I

Test di Mann-Whitney		
Sezione	t_{MW}^{oss}	p-value
Comprensione	27734	0.03
Matematica	30232	<0.01
Logica	28160	0.014
Scienze	32471	<0.01

Tabella 2.6: Statistiche test e p-value dei test di Mann-Whitney per il confronto dei valori mediani tra maschi e femmine nelle sezioni del TOLC_I

e p-value pari a 0.02, il test χ^2 porta al rifiuto dell'ipotesi di uguaglianza tra le due proporzioni, seppur non con forte evidenza. In particolare, la proporzione di OFA conseguiti è significativamente maggiore nel gruppo delle studentesse.

Per quanto riguarda il rendimento universitario, il test di Mann-Whitney porta all'accettazione dell'ipotesi di uguaglianza delle distribuzioni di maschi e femmine per quanto riguarda la variabile `aritmetica` (t_{MW}^{oss}=24868 e p-value=0.967), ma la distribuzione di `numeroesami` risulta significativamente diversa nei due gruppi definiti da `sesso` (t_{MW}^{oss}= 22064 e p-value=0.04). In particolare, la distribuzione del numero di esami sostenuti è spostata verso valori più alti per quanto riguarda le femmine. Il 57.38% delle studentesse ha completato il percorso di studi, tale percentuale è pari al 57,80% per gli studenti.
Il test Chi-quadrato fornisce forte evidenza a favore dell'ipotesi di uguaglianza tra le due proporzioni ($\chi^2 \approx 0$ e p-value=1).

Analisi per corso di laurea

Dopo aver valutato le differenze e le analogie tra studenti in base al sesso, vengono condotte analisi volte a confrontare studenti immatricolati a corsi di laurea diversi.
Il primo confronto riguarda il punteggio conseguito nel test TOLC_I. In Tabella 2.7 sono riportate le principali statistiche di sintesi della variabile `tolc` nei diversi corsi. In Figura 2.17 sono raffigurate le distribuzioni dei punteggi ottenuti al test per corso di laurea.
L'ipotesi di normalità viene rifiutata per quanto riguarda i punteggi degli studenti immatricolati a Fisica (W_{SW}^{oss}=0.975, p-value <0.01) e a Matematica (W_{SW}^{oss}=0.969, p-value <0.01), vengono quindi confrontati i valori mediani delle distribuzioni mediante il test non parametrico di Kruskal-Wallis. Il test porta al rifiuto dell'ipotesi di omogeneità con una statistica test pari a 242.86 (p-value <0.01). In seguito al rifiuto di tale ipotesi, si procede con le analisi post-hoc per identificare quali coppie di mediane sono significativamente diverse tra loro.
Avendo utilizzato il test non parametrico di Kruskal-Wallis, le analisi post-hoc utilizzano la correzione di Holm. Osservando i p-value in Tabella 2.8 si conclude che risultano significativi tutti i confronti tranne quello tra i punteggi degli studenti di Chimica Industriale e Scienze Naturali; in altri termini, si può affermare che le distribuzioni dei punteggi sono tra loro uguali solo per questo confronto.

Corso	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	S.d.	IQR
CHIMICA INDUSTRIALE	8.53	21.11	24.60	24.70	28.48	34.95	5.03	7.37
FISICA	22.77	32.33	34.27	34.42	36.88	41.00	3.18	3.67
MATEMATICA	11.43	26.27	30.80	30.16	34.65	41.00	5.96	8.38
SCIENZE NATURALI	14.15	20.57	23.35	23.77	26.94	34.23	4.81	6.37
STAM	9.03	17.09	22.03	21.01	24.03	35.15	5.09	6.94

Tabella 2.7: Principali statistiche di sintesi delle distribuzioni di `tolc` nei gruppi di `corso`

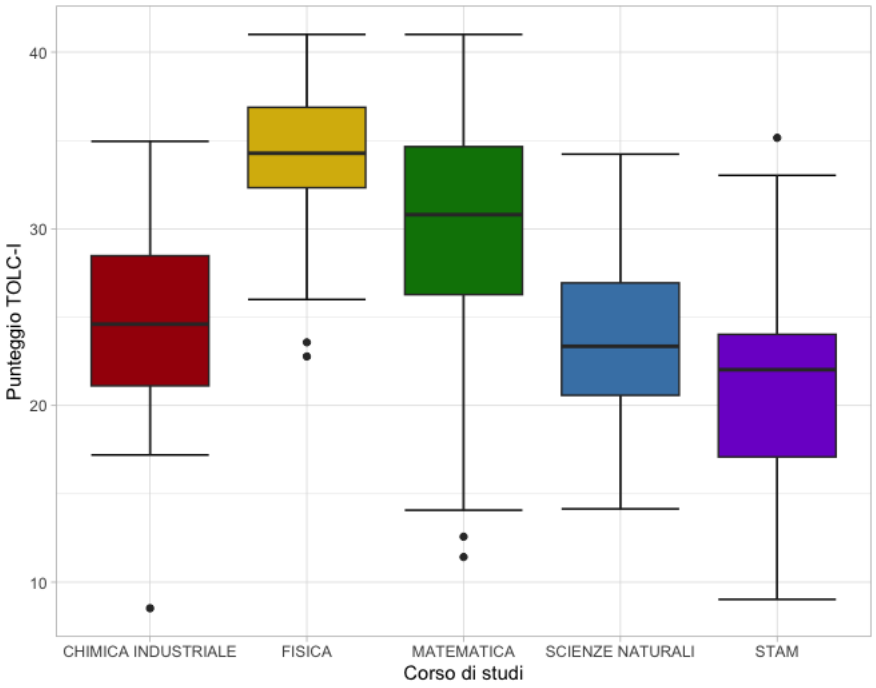

Figura 2.17: Boxplot della variabile `tolc`, per `corso`

Le distribuzioni dei punteggi ottenuti nelle sezioni del test per corso di laurea sono riportate in Figura 2.18. Le distribuzioni dei punteggi degli studenti di Fisica sono spostate verso valori più alti rispetto alle altre, infatti il test di Kruskal-Wallis porta al rifiuto dell'ipotesi di omogeneità per tutte le sezioni. L'evidenza contro tale ipotesi è molto forte per quanto riguarda tutte le sezioni: i p-value dei test di Kruskal-Wallis sono <0.01 in tutti i casi.
Sono state svolte le analisi post-hoc per stabilire, per ogni sezione del test, quali coppie portassero al rifiuto dell'ipotesi di omogeneità. Non vengono riportati nell'elaborato tutti i p-value, ma solo quelli dei confronti che hanno portato a conclusioni interessanti. In seguito a tali analisi, si conclude che i punteggi degli immatricolati a Fisica sono significativamente diversi da quelli degli immatricolati agli altri corsi. In particolare, il rendimento degli studenti di Fisica è migliore degli altri in ogni sezione, seguito da quello degli studenti di Matematica. Nella sezione di Comprensione, l'unica distribuzione significativamente diversa dalle altre è quella degli studenti di Fisica, nessuno degli altri confronti risulta significativo. Nella sezione di Matematica tutti i confronti sono significativi, fatta eccezione per quello tra Chimica Industriale e Scienze Naturali (p-value=0.164). Non c'è evidenza di differenze in distribuzione significative tra gli studenti di Chimica Industriale, STAM e Scienze Naturali per quanto riguarda la sezione di Logica. Infine, il punteggio ottenuto nella sezione di Scienze risulta significativamente diverso per tutti i confronti, fatta eccezione per quello tra Scienze Naturali e STAM.

	CHIMICA IND.	FISICA	MATEMATICA	SCIENZE NATURALI
FISICA	<0.01	–	–	
MATEMATICA	<0.01	<0.01	–	–
SCIENZE NATURALI	0.275	<0.01	<0.01	–
STAM	<0.01	<0.01	<0.01	0.0143

Tabella 2.8: p-value dei test di Mann-Whitney con correzione di Holm per le analisi post-hoc

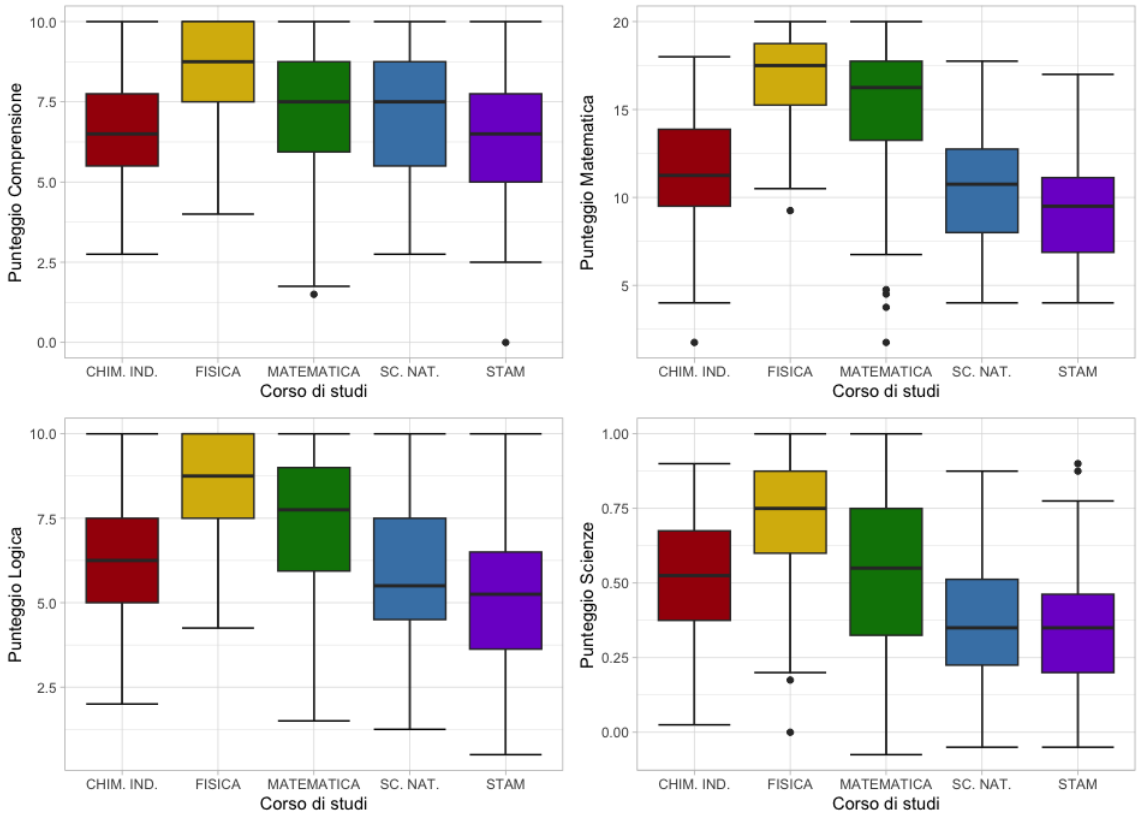

Figura 2.18: Boxplot della variabile **a)** `comprensione`, per `corso`, **b)** `matematica`, per `corso` **c)** `logica`, per `corso` **d)** `scienze`, per `corso`

Le frequenze assolute e percentuali di studenti che hanno conseguito l'OFA in Matematica nei 5 corsi sono riportate in Tabella 2.9. Si nota una percentuale di OFA elevata tra gli Studenti di STAM: quasi un quarto degli iscritti lo ha conseguito. Tra i 165 iscritti al corso di Fisica, nessuno ha conseguito l'obbligo formativo aggiuntivo.

La distribuzione del numero di esami sostenuti (`numeroesami`) risulta significativamente diversa a seconda del corso di laurea: il test di Kruskal-Wallis porta al rifiuto dell'ipotesi di omogeneità con un p-value<0.01. Le analisi post-hoc portano a concludere che la distribuzione di `numeroesami` per gli studenti di Fisica è significativamente diversa da quelle di tutti gli altri corsi tranne Matematica.

Per quanto riguarda la media aritmetica, il test di Kruskal-Wallis porta al rifiuto dell'ipotesi di omogeneità (p-value<0.01). In seguito alle analisi post-hoc si conclude che la differenza in distribuzione tra gli studenti di Fisica e quelli di Matematica porta al rifiuto dell'ipotesi di omogeneità: il p-value del test di Mann-Whitney per tale confronto è <0.01, tutti gli altri confronti non risultano significativi (p-value>0.05 per ogni confronto).

	CHIMICA IND.	FISICA	MATEMATICA	SCIENZE NATURALI	STAM
Assolute	1	0	4	5	14
Percentuali	2.1%	0.0%	3.0%	9.1%	23.7%

Tabella 2.9: Frequenza percentuale di studenti che hanno conseguito OFA, per `corso`

La percentuale di laureati dopo 3 anni è minore del 50% nei corsi di Fisica e Matematica. In Tabella 2.10 sono riportate alcune statistiche riguardanti la frequenza di laureati in ogni corso e il loro rendimento in termini di voto medio di laurea. La distribuzione della variabile `votolaurea` è significativamente diversa nei 5 gruppi (p-value del test di Kruskal-Wallis <0.01). Attraverso le analisi post-hoc si conclude che l'unica differenza fortemente significativa in termini di voto di laurea si ha tra Fisica e STAM (p-value<0.01).

Corso	Laureati	% Laureati	Voto medio (S.d.)	Mediana
CHIMICA INDUSTRIALE	25	53.2%	101.60 (6.09)	102.00
FISICA	64	38.8%	105.86 (4.70)	107.50
MATEMATICA	41	31.06%	103.07 (5.61)	104.00
SCIENZE NATURALI	29	52.7%	105.38 (5.60)	108.00
STAM	35	59.3%	101.23 (6.45)	102.00

Tabella 2.10: Statistiche di sintesti delle variabili `laureato` e `votolaurea`, per `corso`

Analisi tra variabili quantitative

In questa ultima sezione vengono analizzate le relazioni tra le variabili quantitative attraverso diagrammi di dispersione e indici di correlazione. È di particolare interesse la valutazione della correlazione tra il voto del TOLC_I e la media aritmetica degli esami sostenuti. La scelta di studiare la relazione tra voto del test (`tolc`) e media aritmetica (variabile `aritmetica`) è dovuta al fatto che tale informazione è disponibile per tutte le unità statistiche, mentre `ponderata` e `votolaurea` sono disponibili, ovviamente, solo per chi ha completato il percorso di studi. Inoltre, le correlazioni positive quasi perfette tra `aritmetica`, `ponderata` e `votolaurea` fanno sì che il comportamento di queste variabili sia pressochè analogo. Per una rappresentazione grafica di tali correlazioni si veda la Figura 2.19.

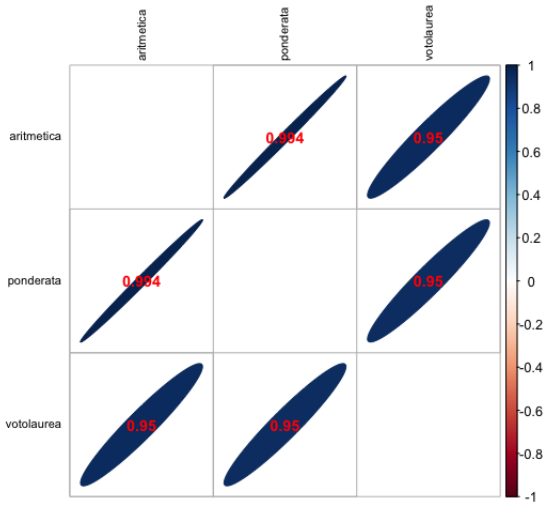

Figura 2.19: Corrplot delle variabili `aritmetica`,`ponderata` e `votolaurea`

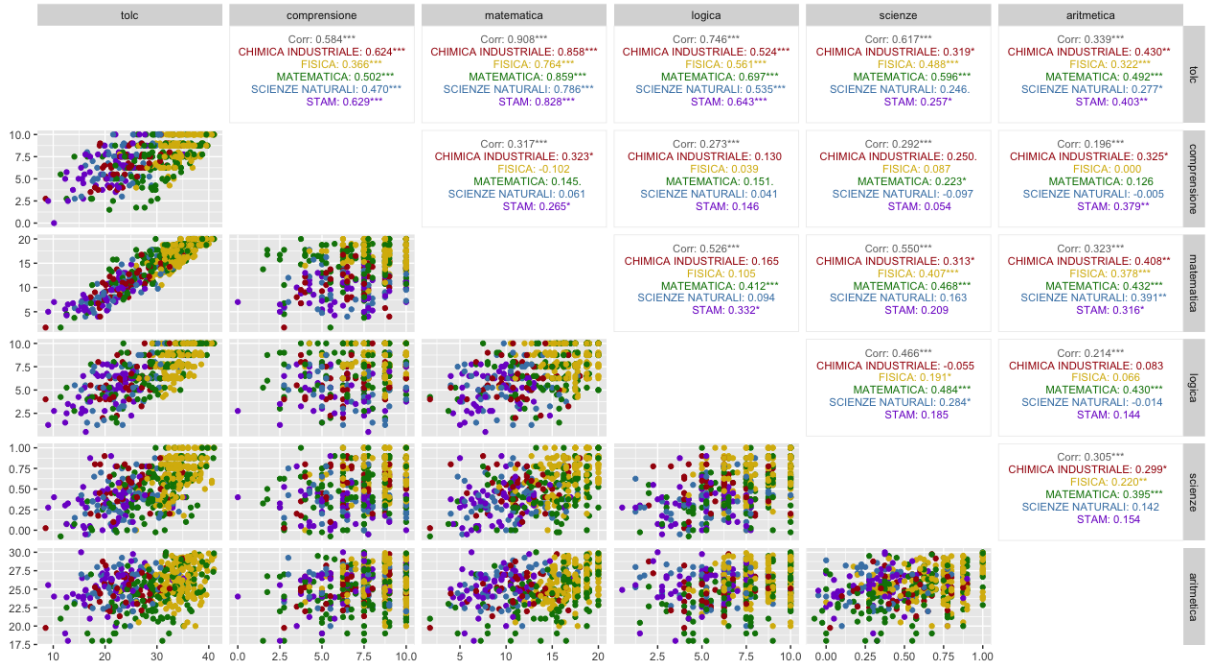

Figura 2.20: Diagrammi di dispersione delle variabili `tolc`, `matematica`, `logica`, `comprensione`, `scienze` e `aritmetica`

La Figura 2.20 racchiude in sè una grande mole di informazione utile allo studio della relazione tra il voto del test e la media aritmetica. Osservando i diagrammi di dispersione e i valori dei coefficienti di correlazione riportati nel grafico si cerca di stabilire se la media aritmetica sia correlata con il punteggio del TOLC_I, o con il punteggio ottenuto in qualche sezione in particolare. La media aritmetica non risulta particolarmente correlata con il punteggio ottenuto in qualche sezione specifica ($r_{xy} < 0.35$ per ogni coppia). Il punteggio ottenuto nella sezione di Matematica è piuttosto correlato al punteggio totale del test ($r_{xy} = 0.908$): questo valore può essere spiegato dal fatto che la sezione di Matematica costituisce quasi la metà del test (20 punti su 41). Inoltre, il grafico in Figura 2.20 conferma quanto concluso nello scorso paragrafo: gli studenti immatricolati a Fisica hanno avuto il rendimento migliore al TOLC_I, sia per quanto riguarda il punteggio generale sia per quanto riguarda le singole sezioni (lo si può notare osservando i diagrammi di dispersione nella parte in basso a sinistra del grafico).
In conclusione, viene confermato il primato degli studenti di Fisica per quanto riguarda il rendimento nel test. Tuttavia non si nota la presenza di una evidente relazione lineare tra il punteggio del TOLC_I (o di qualche sezione in particolare) e il rendimento universitario in termini di media aritmetica.

In Figura 2.21 viene riportato il diagramma di dispersione delle variabili `tolc` e `aritmetica`. Le osservazioni sono distinte per colore secondo le modalità della variabile `laureato`. Il coefficiente di correlazione tra `tolc` e `aritmetica` è pari a 0.36 tuttvia si nota che i punti in corrispondenza degli studenti laureati sono spostati verso valori alti. Per il gruppo dei laureati, l'indice di correlazione tra `tolc` e `aritmetica` è pari a 0.49, mentre per i non laureati è 0.35. Il punteggio ottenuto al TOLC_I è moderatamente correlato al rendimento universitario (rappresentato in termini di media aritmetica) per quanto riguarda gli studenti che hanno completato il percorso di studi, mentre tale correlazione è più debole per i non laureati.

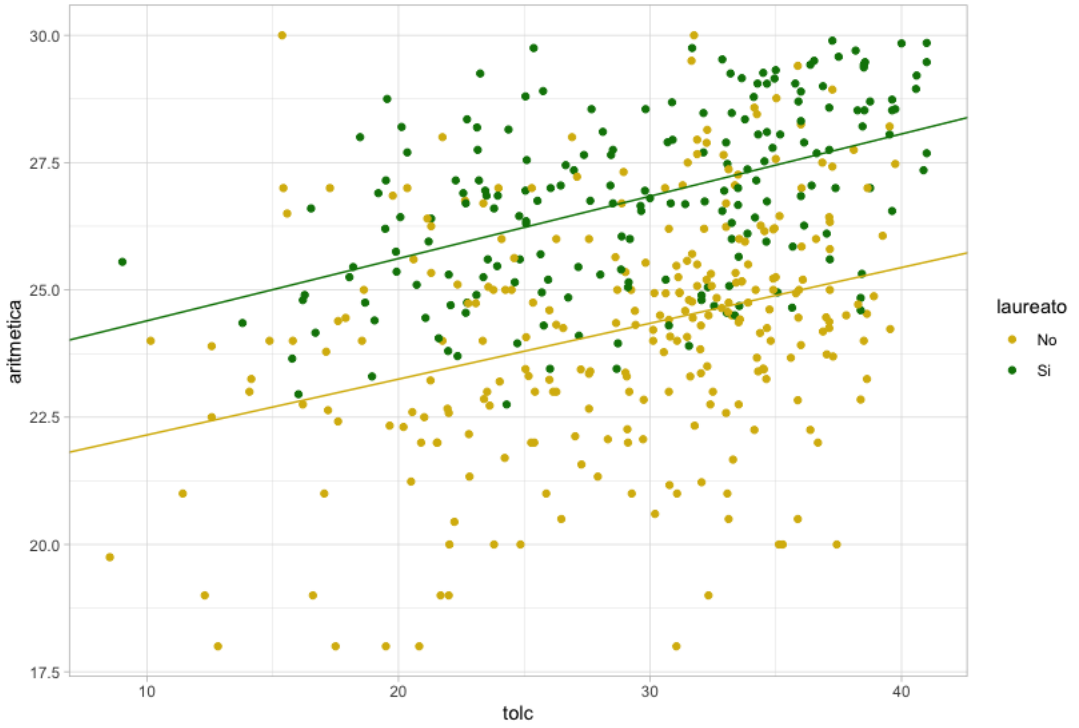

Figura 2.21: Diagramma di dispersione delle variabili `tolc` e `aritmetica`, per `laureato`

Nel prossimo capitolo vengono descritti i comportamenti delle variabili di interesse in funzione delle altre variabili mediante l'implementazione di modelli di regressione diversi a seconda della natura della variabile risposta.

Capitolo 3

Analisi multivariate

Un problema centrale della Statistica è studiare la relazione tra una variabile risposta, o una sua trasformazione, e altre variabili, indicate in genere indistintamente con i termini variabili esplicative, o variabili concomitanti, o predittori. Si utilizza 'variabili concomitanti' con riferimento alle variabili disponibili nell'insieme di dati da analizzare, mentre si riserva il termine 'variabili esplicative' alle variabili, definite a partire da un insieme selezionato di variabili concomitanti, che compaiono nella formula che specifica un modello di regressione [3].
In questo capitolo vengono utilizzate come risposte le variabili `votolaurea`, `numeroesami` e `laureato`. Per ciascuna di queste tre variabili vengono implementati uno o più modelli di regressione che permettono di studiarne la relazione con alcune variabili esplicative. La scelta delle variabili esplicative avviene tramite procedure di selezione in avanti: partendo dal modello con sola intercetta, vengono svolti opportuni test per valutare quali siano le variabili utili a descrivere il comportamento della risposta. Il procedimento di selezione si interrompe quando, dopo aver aggiunto alcune esplicative, tutte le altre variabili risultano non significative.

3.1 Voto laurea

In questa sezione vengono riportati e interpretati i risultati principali dei modelli di regressione aventi come variabile risposta `votolaurea`. In primo luogo, viene studiata la relazione tra `votolaurea` e le varibili esplicative attraverso l'implementazione di un modello di regressione lineare normale. Successivamente, viene adattato un modello Tobit [4] per valutare eventuali miglioramenti. In statistica, un modello Tobit è un qualunque modello di regressione in cui il range osservato per la variabile risposta è limitato in qualche modo [4]. Dal momento che il dominio della variabile `votolaurea` è limitato sia inferiormente sia superiormente (il voto di laurea varia tra 66 e 110), può essere ragionevole adattare un modello di questo tipo.

3.1.1 Modello lineare normale

Il procedimento di selezione in avanti porta a includere nel modello le seguenti variabili esplicative: `tolc`, `corso`, `maturita`, `scienze`. Per il procedimento di selezione sono state rimosse dal dataset di partenza alcune variabili come `aritmetica` e `ponderata`. Ad esempio, mantenendo `ponderata` tra le variabili selezionabili, il procedimento porta a includerla come unica esplicativa nel modello, in grado di spiegare il 90% della variabilità di `votolaurea` ($R^2 = 0.9$). Ciò è dovuto al fatto che `ponderata` e `votolaurea` sono

quasi perfettamente correlate: come visto nel Cap. 2, la correlazione tra queste due variabili è prossima a 1. In Tabella 3.1 vengono riportate le stime dei coefficienti e i relativi standard error, oltre alle statistiche test e i relativi p-value per verificare la significatività delle stime.

Coefficienti	Stima	I.C. 95%	Std. Error	t value	$Pr(>\|t\|)$
Intercetta	71.21	[62.45; 79.98]	4.44	16.03	< 0.01
tolc	0.32	[0.13; 0.51]	0.09	3.40	< 0.01
corsoFISICA	-2.01	[-5.00; 0.98]	1.52	-1.32	0.19
corsoMATEMATICA	-3.28	[-6.18; -0.38]	1.47	-2.23	0.03
corsoSCIENZE NATURALI	6.37	[3.57; 9.17]	1.42	4.49	< 0.01
corsoSTAM	3.21	[0.52; 5.90]	1.36	2.36	0.02
maturita	0.22	[0.14; 0.31]	0.04	5.18	< 0.01
scienze	4.82	[1.35; 8.29]	1.75	2.74	< 0.01

Tabella 3.1: Stime del modello lineare normale per `votolaurea`

Tutti i coefficienti, fatta eccezione per il coefficiente relativo alla modalità 'Fisica' della variabile `corso`, risultano significativi al 5%. Nel complesso si può confermare la significatività di `corso`, data la significatività dei coefficienti delle altre modalità di tale variabile e dato il fatto che il test F per il confronto tra modelli annidati porta a preferire il modello che include tale variabile (p-value< 0.01).
I coefficienti per `tolc`, `maturita` e `scienze` vengono interpretati come variazione in termini di voto di laurea medio per un aumento unitario della variabile a cui fanno riferimento, a parità di tutte le altre esplicative. Nello specifico, l'aumento di un punto nel punteggio del test TOLC_I comporta un aumento di 0.32 punti nel voto medio di laurea. Analogamente, l'aumento di un voto all'esame di maturità è associato all'aumento di 0.22 punti sul voto di laurea medio. Infine, l'aumento di un punto nella sezione di scienze comporta un aumento di 4.82 punti sul voto di laurea. Per la variabile `corso` l'intepretazione è leggermente diversa. Trattandosi di una variabile qualitativa, nel modello compaiono un numero di coefficienti pari al numero di modalità della variabile meno uno. La modalità a cui non corrisponde alcun coefficiente è detta 'di riferimento'. In questo caso, la modalità di riferimento è 'Chimica Industiale'. Il coefficiente di 'Matematica' pari a -3.28 indica che il voto di laurea medio per gli studenti iscritti a Matematica è minore di 3.28 punti rispetto a quello per gli studenti di Chimica Industriale. Allo stesso modo, il voto medio per per gli studenti di Scienze Naturali e STAM è, rispettivamente, maggiore di 6.37 e 3.21 punti rispetto agli studenti di Chimica Industriale. Non viene riportata l'interpretazione del coefficiente per gli studenti di Fisica in quanto non risulta signficativo. Si noti che anche nel caso di variabili qualititative, l'interpretazione dei coefficienti è da intendersi come variazione in termini di valore medio della risposta rispetto alla modalità di riferimento a parità di tutte le altre variabili.

Bontà di adattamento

L'indice R^2 pari a 0.32 indica che il modello spiega il 32% della variabilità del voto di laurea. In altre parole, le variabili esplicative incluse nel modello riescono a spiegare il 32% della variabilità di `votolaurea`. Il criterio di informazione di Akaike (AIC) è pari a 1175.27. Tale indice è utile per confrontare modelli di regressione di diverso tipo. Un altro indice per valutare la bontà del modello è l'indice di correlazione multipla indicato con R. Tale indice misura la correlazione tra i valori osservati della risposta e quelli stimati sotto il modello. Per questo modello, si ha $R = 0.57$. Insieme all'AIC, questo indice sarà usato per il confronto con il modello Tobit.

3.1.2 Modello Tobit

Il modello di regressione lineare normale tratta i valori di `votolaurea` a 110 come valori qualunque e non come valore massimo del voto di laurea [4]. Per questo motivo viene adattato un modello Tobit che tenga conto del dominio limitato di `votolaurea`.
Le variabili esplicative incluse nel modello sono le stesse del modello lineare normale. L'AIC del modello pari a 1006.26, è inferiore a quello del modello lineare normale e suggerisce che il modello Tobit sia preferibile rispetto all'altro. In Tabella 3.2 sono riportate le stime dei coefficienti, i relativi standard error, le statistiche test e i relativi p-value ottenuti con il modello Tobit.

Coefficienti	Stima	I.C. 95%	Std. Error	z value	$Pr(>\|z\|)$
Intercetta:1	62.41	[50.91; 73.91]	5.87	10.64	< 0.01
Intercetta:2	1.82	[1.71; 1.94]	0.06	29.88	< 0.01
tolc	0.49	[0.24; 0.73]	0.13	3.89	< 0.01
corsoFISICA	-3.15	[-7.08; 0.77]	2.00	-1.57	0.12
corsoMATEMATICA	-4.63	[-8.40; -0.86]	1.93	-2.41	0.02
corsoSCIENZE NATURALI	7.96	[4.28; 11.64]	1.88	4.24	< 0.01
corsoSTAM	4.18	[0.71; 7.65]	1.77	2.363	0.02
maturita	0.28	[0.17; 0.39]	0.06	4.94	< 0.01
scienze	5.44	[0.88; 9.99]	2.32	2.34	0.02

Tabella 3.2: Stime del modello Tobit per `votolaurea`

Le interpretazioni dei coefficienti di questo modello sono analoghe a quelle del modello lineare normale, con la differenza che l'effetto lineare è sulla variabile latente non censurata, non sulla risposta osservata [4]. Rispetto al modello lineare, le stime dei coefficienti sono leggermente più alte in valore assoluto. Il coefficiente riguardante il corso di Fisica è ancora non significativo. Il coefficiente per `maturita` è molto vicino a quello stimato col primo modello (0.28 per il modello Tobit, 0.22 per il modello lineare normale). Per quanto riguarda `tolc`, un incremento di un punto sul voto ottenuto al test comporta un aumento di mezzo punto (0.49) sulla stima voto di laurea. Il voto di laurea stimato sotto il modello per gli studenti di Matematica è minore di 4.63 punti rispetto a quelli di Chimica Industriale, mentre per gli studenti di Scienze Naturali e STAM è rispettivamente maggiore di 7.96 e 4.18 punti.

Bontà di adattamento

L'indice R è pari a 0.57, come nel modello lineare normale. Entrambi i modelli risultano complessivamente validi, ma sulla base dell'AIC il modello Tobit sembra preferibile rispetto al modello lineare normale. Inoltre, data la natura di `votolaurea`, un modello Tobit risulta più appropriato grazie alla sua capacità di tenere conto del dominio limitato.

3.2 Numero esami

Dopo aver indagato la relazione tra `votolaurea` e le altre variabili del dataset, viene studiato il comportamento di `numeroesami` mediante l'implementazione di tre modelli di regressione. Di seguito vengono riportati solo due dei tre modelli implementati, ovvero il modello lineare normale e un modello per dati di conteggio che tiene conto della sovradispersione (modello binomiale negativo) [3].

È stato implementato anche un modello Tobit, ma non viene riportato a causa dell'indice R particolarmente basso rispetto agli altri due modelli ($R = 0.387$).

3.2.1 Modello lineare normale

Al termine del procedimento di selezione delle variabili, il modello lineare normale per `numeroesami` include come esplicative `aritmetica`, `laureato`, `matematica` e `sesso`. In Tabella 3.3 viene riportata una sintesi delle stime ottenute con questo modello.

Coefficienti	Stima	I.C. 95%	Std. Error	t value	Pr($>\|t\|$)
Intercetta	-6.50	[-11.66; -1.33]	2.63	-2.47	0.01
aritmetica	0.61	[0.39; 0.84]	0.11	5.37	< 0.01
laureatoSi	7.95	[6.88; 9.01]	0.54	14.69	< 0.01
matematica	0.17	[0.06; 0.28]	0.06	2.95	< 0.01
sessoMASCHIO	-1.21	[-2.15; -0.28]	0.48	-2.54	0.01

Tabella 3.3: Stime del modello lineare normale per `numeroesami`

Tutte le stime sono significative e si ricorda che le loro interpretazioni sono da intendersi come effetto sulla media della variabile risposta a parità di tutte le altre variabili.
Il coefficiente per `laureato` pari a 7.95 indica che uno studente laureato sostiene mediamente quasi 8 esami in più rispetto a uno non laureato. Per quanto riguarda il sesso, i maschi sostengono mediamente 1.21 esami in meno rispetto alle femmine. Le variabili `aritmetica` e `matematica` sono positivamente associate a `numeroesami`, come dimostrano le stime dei coefficienti maggiori di zero. In particolare, ad un aumento di un punto della media artimetica corrisponde un aumento medio di 0.61 esami sostenuti. Anche il voto ottenuto nella sezione di matematica del TOLC_I ha un effetto significativo su `numeroesami`: per un aumento di un punto in tale sezione si ha un incremento medio di 0.17 esami sostenuti. Si può quindi affermare che all'aumentare della media aritmetica corrisponde un aumento del numero medio esami sostenuti; la stessa relazione, seppur in misura leggermente ridotta, si ha tra il punteggio nella sezione di matematica e la variabile risposta.

Bontà di adattamento

L'AIC del modello è pari a 2755.13 e sarà utile per il confronto con il modello per dati di conteggio. Per quanto riguarda l'indice di correlazione multipla si ha $R = 0.72$. Il modello è in grado di spiegare il 52.3% della variabilità di `numeroesami` ($R^2 = 0.523$).

3.2.2 Modello binomiale negativo

Dal momento che i valori della risposta rappresentano il risultato di conteggi il cui totale non è prefissato, il modello di riferimento per la distribuzione di `numeroesami` è il modello di Poisson [3]. Tuttavia, in seguito all'implementazione di un modello di Poisson per `numeroesami` emerge che le assunzioni alla base del modello non sono rispettate. La violazione di tali assunzioni suggerisce la presenza di sovradispersione nei dati. È stato quindi implementato un modello che possa tenere conto della sovradispersione, ovvero il modello binomiale negativo. Per approfondimenti su dati di conteggio e sovradispersione si veda Salvan [3]. La selezione in avanti ha portato a includere nel modello di Poisson 5 variabili esplicative:

`aritmetica`, `laureato`, `matematica`, `sesso` e `corso`. Nel modello binomiale negativo, la variabile `corso` non risulta più significativa. In Tabella 3.4 vengono riportate le stime ottenute con il modello binomiale negativo.

Coefficienti	Stima	I.C. 95%	Std. Error	z value	$\Pr(>\|z\|)$	$exp(\hat{\beta})$
Intercetta	0.46	[-0.11; 1.04]	0.27	1.71	0.09	1.59
aritmetica	0.07	[0.04; 0.09]	0.01	5.83	< 0.01	1.07
laureatoSi	0.53	[0.43; 0.63]	0.05	9.91	< 0.01	1.70
matematica	0.02	[0.00; 0.03]	0.01	2.68	< 0.01	1.02
sessoMASCHIO	-0.10	[-0.20; -0.01]	0.05	-2.19	0.03	0.90

Tabella 3.4: Stime del modello binomiale negativo per `numeroesami`

La presenza di sovradispersione è confermata in quanto la stima del parametro τ, che regola la sovradispersione, è maggiore di zero. Inoltre, l'intervallo di confidenza al 95% per tale parametro non include lo zero ($\tau = 0.16$, I.C. 95%: [0.13;0.21]). Il modello è stato implementato utilizzando la funzione di legame canonica (funzione logaritmica), perciò per la corretta interpretazione dei coefficienti bisogna considerare le quantità $exp(\beta_i)$ per $i = 1, ..., 5$, anch'esse riportate nella Tabella 3.4.

Per un aumento unitario della media aritmetica, il numero medio di esami sostenuti aumenta di 1.07 volte, mentre per un aumento unitario del punteggio conseguito nella sezione di matematica il numero medio di esami sostenuti aumenta di 1.02 volte. Per quanto riguarda queste variabili viene in parte confermato quanto emerso dal modello lineare normale: l'effetto di `aritmetica` sulla variabile risposta è leggermente più marcato di quello di `matematica`, anche se in generale l'effetto di queste due variabili sulla risposta non è molto forte. Per quanto riguarda le variabili qualitative, il numero medio di esami sostenuti da un soggetto laureato è 1.70 volte tale numero per uno studente non laureato. Per uno studente maschio, il numero medio di esami sostenuti è 0.9 volte quello di una studentessa.

Bontà di adattamento

L'AIC del modello è pari a 3055, più alto di quello del modello lineare normale. Anche l'indice R porta a preferire il primo modello: per il modello binomiale negativo tale indice è pari a 0.68, leggermente minore rispetto al modello lineare normale.

3.3 Laureato

L'ultimo modello considerato è quello che ha come risposta la variabile dicotomica `laureato`. In questa sezione viene riportato un solo modello logistico che permette di modellare la probabilità di laurearsi. Dal dataset di partenza viene rimossa la variabile `numeroesami` dal momento che in corrispondenza di studenti laureati `numeroesami` assume necessariamente valore massimo.
Nella parte conclusiva di questa sezione viene commentato un modello ottenuto mantenendo `numeroesami` tra le concomitanti, ma senza riportarne tutti i risultati.

3.3.1 Modello logistico senza numero esami

Il procedimento di selezione in avanti porta a definire un modello che ha come esplicative `aritmetica` e `corso`.

Come si può vedere in Tabella 3.5, tutte le variabili risultano significative. Per quanto riguarda `corso` possiamo confermare la significatività complessiva vista l'alta significatività delle modalità Matematica e Fisica. Il modello è stato implementato utilizzando la funzione di legame canonica (logit) e le stime dei coefficienti rappresentano i logaritmi degli odds ratio. Per approfondimenti su modelli logistici, funzione di legame e odds ratio si veda, ad esempio, Salvan *et al.* (2020, Capitolo 3). Anche in questo caso non vengono interpretate direttamente le stime $\hat{\beta}$ dei coefficienti, ma le quantità $exp(\hat{\beta})$, che costituiscono i rapporti tra quote (*odds ratio*), utili per confrontare le probabilità di laurearsi in diverse circostanze. Per variabili qualitative, $exp(\hat{\beta})$ rappresenta la variazione percentuale della probabilità di laurearsi corrispondente a un aumento unitario della variabile esplicativa a cui il coefficiente β fa riferimento, fermo restando il valore delle ulteriori esplicative del modello. In caso di variabili dicotomiche, $exp(\hat{\beta})$ rappresenta la variazione moltiplicativa della probabilità di laurearsi quando $x = 1$ rispetto a quando $x = 0$, mantendo costanti i valori delle altre variabili esplicative. Tali quantità sono riportate in Tabella 3.5.

Coefficienti	Stima	I.C. 95%	Std. Error	z value	$\Pr(>\|z\|)$	$exp(\hat{\beta})$
Intercetta	-16.58	[-20.10; -13.35]	1.72	-9.65	< 0.01	0.00
aritmetica	0.67	[0.54; 0.81]	0.07	9.90	< 0.01	1.95
corsoFISICA	-1.47	[-2.28; -0.68]	0.41	-3.36	< 0.01	0.22
corsoMATEMATICA	-1.01	[-1.83; -0.10]	0.42	-2.43	0.02	0.36
corsoSCIENZE NATURALI	-0.30	[-1.24; 0.63]	0.48	-0.64	0.53	0.74
corsoSTAM	0.14	[-0.77; -1.06]	0.46	0.31	0.76	0.87

Tabella 3.5: Stime del modello logistico per `laureato` ed effetti delle variabili

Di seguito vengono riportate solo le interpretazioni delle stime significative. A parità delle altre variabili, per un aumento unitario della media aritmetica, la quota di laureati varia di 1.95 volte.
In altre parole, ad un aumento di un punto della media aritmetica corrisponde un incremento del 95% della quota di laureati. Per quanto riguarda `corso`, la modalità di riferimento è ancora Chimica Industriale. La quota stimata per la probabilità che uno studente sia laureato per gli studenti di Fisica e Matematica è rispettivamente 0.22 e 0.36 volte la stessa quota per gli studenti di Chimica Industriale. Ciò signfica che la stima della probabilità di laurearsi entro tre anni per gli studenti di Fisica è minore dell'78% rispetto a quelli di Chimica Industriale, mentre per gli studenti di Matematica è minore del 64% rispetto alla modalità di riferimento.

Bontà di adattamento

Uno degli strumenti maggiormente utilizzati per valutare la bontà di adattamento del modello è la curva ROC (*Receiver Operating Characteristic*), che fornisce una misura della capacità predittiva del modello. Un indicatore di sintesi della curva ROC è l'AUC (*Area Under the Roc Curve*): tale indicatore varia tra 0 e 1 e quanto più si avvicina a 1, tanto migliore è la capacità predittiva del modello. In Figura 3.1 viene riportata la curva ROC. L'AUC pari a 0.839 indica una buona capacità predittiva per il modello implementato.
Un ulteriore strumento per valutare la bontà di adattamento di un modello logistico è la statistica di Hosmer e Lemeshow. Per il modello implementato, il valore di tale statistica è 21.77 con relativo p-value pari a 0.06. L'adattamento del modello ai dati è quindi soddisfacente.

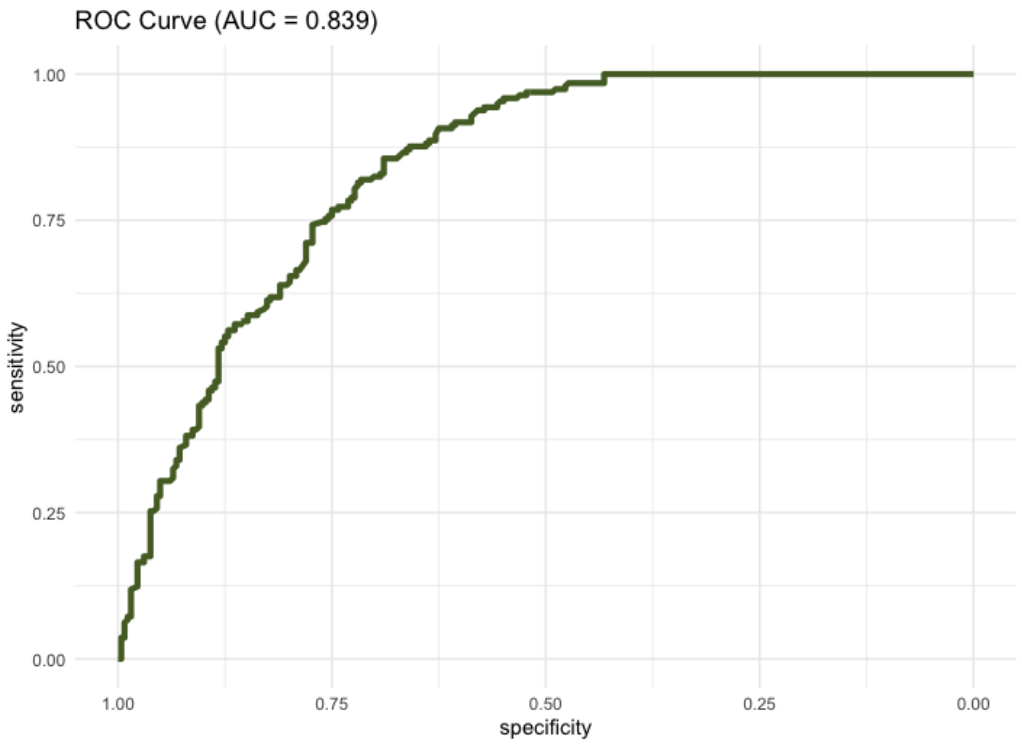

Figura 3.1: Curva ROC per il modello logistico

3.3.2 Modello logistico con numero esami

Includendo `numeroesami` tra le variabili selezionabili per far parte del modello, essa entrerebbe a far parte del modello al primo passo della selezione in avanti. Le variabili esplicative di tale modello sarebbero, oltre a `numeroesami`, `aritmetica` e `sesso`. Senza entrare nei dettagli delle stime, il modello sarebbe caratterizzato da un AUC particolarmente alto (0.95) che indica una capacità predittiva quasi perfetta. Bisogna però tenere in considerazione il fatto che tale valore per l'AUC è dovuto alla relazione deterministica tra `numeroesami` e la risposta `laureato`. Infatti, a meno di studenti che hanno sostenuto tutti gli esami ma non hanno ancora conseguito la laurea, il valore massimo di `numeroesami` per uno studente implica che sia laureato. Per questo motivo, nonostante l'elevatissima capacità predittiva del modello e nonostante il valore del Criterio di Informazione di Akaike sia di gran lunga inferiore a quello del primo modello (245.80 contro 470.46), il modello che non include `numeroesami` tra le esplicative è da considerare come migliore.

Capitolo 4

Conclusioni

Ripercorrendo le analisi si conclude che il punteggio del test conseguito dagli studenti è significativamente maggiore rispetto a quello delle studentesse. Tale differenza appare piuttosto marcata specialmente nelle sezioni di Matematica e Scienze, dove l'evidenza a favore dell'ipotesi di disomogeneità delle distribuzioni è molto forte (p-value<0.01). Per quanto riguarda il corso di laurea, si osservano i punteggi migliori in corrispondenza degli studenti di Fisica, mentre i più bassi sono quelli degli immatricolati al corso di STAM. I punteggi degli studenti di Matematica appaiono piuttosto variabili, ciò può essere giustificato dal fatto che è un corso ad accesso libero, al contrario degli altri 4 che sono ad accesso programmato e richiedono, almeno per la selezione anticipata del periodo primavera-estate, il conseguimento di un punteggio minimo al test. Anche all'interno delle sezioni del test l'andamento rispecchia quanto osservato per il punteggio generale, confermando il primato degli studenti di Fisica.
Il voto di laurea medio per il corso di Fisica (105.86) è il più alto tra i 5 corsi, il più basso è quello di STAM (101.23). Per quanto riguarda la percentuale di laureati, invece, si registra la frequenza più alta proprio a STAM (59.3%), mentre a Fisica la seconda più bassa (38.8%). In conclusione gli studenti di Fisica si laureano con voti mediamente migliori, ma la percentuale di laureati dopo 3 anni è tra le più basse nel campione. È importante precisare che i dati non tengono conto degli studenti che abbandonano l'università o cambiano corso di laurea, quindi la percentuale di laureati è calcolata sul numero di immatricolati al primo anno. La percentuale dei laureati nel corso di Fisica è molto influenzata da questo fenomeno. Al contrario, gli studenti di STAM ci mettono meno tempo a laurearsi, ma lo fanno conseguendo voti mediamente più bassi rispetto agli studenti degli altri corsi.
Per quanto riguarda la relazione tra il punteggio del TOLC_I e il rendimento universitario, si osserva un coefficiente di correlazione positivo seppur non particolarmente alto (0.34). Quindi all'aumentare del voto conseguito nel test, il rendimento migliora. Si ricordi che il rendimento universitario è stato spesso espresso in termini di media aritmetica dal momento che l'alta correlazione con le altre variabili rappresentanti il rendimento (`ponderata` e `votolaurea`) indica forti analogie in comportamento tra di esse. La forza della relazione lineare tra il voto del test e il rendimento universitario è più forte per i laureati piuttosto che per i non laureati.
Il voto conseguito nel test risulta avere un effetto significativo sul voto di laurea. Oltre al punteggio generale, anche il punteggio della sezione di scienze influisce sul voto di laurea. Il voto medio di laurea è influenzato anche dal voto di maturità, anche se con un'intensità minore rispetto al tolc. Per quanto riguarda il numero medio di esami sostenuti, non si registra un effetto del punteggio generale del test, ma il coefficiente relativo al punteggio della sezione di matematica risulta significativo. Si può dunque affermare che il rendimento nel test influenza il numero di esami sostenuti, data l'elevata correlazione tra

`matematica` e `tolc` ($r_{xy} = 0.908$). Sulla probabilità di laurearsi, invece, non vi è evidenza dell'influenza del punteggio ottenuto nel test. Media aritmetica e corso di laurea sono sufficienti per definire un modello con una discreta bontà di adattamento (AUC=0.839).
Si conclude quindi che c'è un effetto del test sul rendimento universitario, in particolare sul numero medio di esami sostenuti e sul voto di laurea. Tuttavia, tale rendimento appare influenzato anche da altri fattori, specialmente dal corso di laurea. Infine, non ci sono evidenze di differenze significative tra il rendimento di maschi e femmine se non per quanto riguarda il numero di esami sostenuti.

Appendice

In questa sezione viene spiegato come è stato possibile ricavare un dataset che raccogliesse per riga i voti degli esami sostenuti da ciascuno studente utilizzando il codice fiscale (`cf` nel codice) come identificatore univoco dello studente. I dati iniziali erano organizzati in un dataset contenente una riga per ogni esame sostenuto da un singolo studente. Ciò significa che ogni studente figurava nel dataset un numero di volte pari al numero di esami da esso sostenuti. Una volta importati i dati sul software SAS ed una volta individuato il numero massimo di esami sostenuti da un singolo studente (21), è stato utilizzato il codice che segue per la creazione del dataset nel formato desiderato.

Codice SAS per la riorganizzazione del dataset sugli esami

```
PROC SORT DATA=lib.esamicoorti2019;
BY cf;
RUN;

DATA tuttigliesamisuunariga (keep=cf votoesame1-votoesame21);
RETAIN cf votoesame1-votoesame21;
ARRAY avoti (max) votoesame1-votoesame21;

SET lib.esamicoorti2019;
BY cf;
IF FIRST.cf THEN DO;
     i=1;
     DO j=1 TO max;
          avoti(j)=.;
     END;
END;
avoti(i)=votoesame;

IF LAST.cf THEN OUTPUT;
i+1;
RUN;
```

A questo punto disponiamo di un dataset chiamato `tuttigliesamisuunariga` organizzato nel modo desiderato. Questo dataset è stato esportato in excel ed è stata aggiunta una colonna `controlloesami` con valore 1 per ogni riga: servirà alla fine del procedimento per garantire che in corrispondenza degli

studenti inclusi nel campione siano a disposizione le informazioni sugli esami sostenuti all'università. Con poche righe di codice si effettua il merging dei tre datasets. Per comodità indichiamo i datasets nel codice con i seguenti nomi:

- `tolc`: dataset contenente le informazioni sui punteggi conseguiti al TOLC_I e le informazioni sul percorso di istruzione superiore. Anche in questo dataset è presente una colonna `controllotolc` allo stesso scopo di quella inserita nel dataset sugli esami.
- `esami`: il dataset che avevamo chiamato `tuttigliesamisuunariga`.
- `laureati`: dataset che contiene medie aritmetiche e ponderate, voti di laurea ed eventuale lode per ciascuno studente.

Codice SAS per il merging dei datasets

```
PROC SORT DATA=tolc;
BY CF;
RUN;

PROC SORT DATA=esami;
BY CF;
RUN;

PROC SORT DATA=laureati;
BY CF;
RUN;

DATA tolcesami;
MERGE tolc esami;
BY CF;
RUN;

DATA datifinali;
MERGE tolcesami laureati;
BY CF;
RUN;
```

Il dataset `datifinali` contiene tutti i dati necessari per svolgere le analisi. Sono presenti le colonne rappresentanti le variabili `controllotolc` e `controlloesami`. Come ultimo passo, il dataset viene esportato in excel e viene aggiunta la colonna `controllototale`: i valori di questa variabile sono dati dalla somma dei valori di `controllotolc` e `controlloesami`. Gli studenti per cui la variabile `controllototale` ha valore 2 sono quelli per cui si hanno a disposizione sia i dati sull'esito del TOLC_I sia quelli sugli esami universitari e che quindi compongono il campione da analizzare.

Elenco delle figure

Elenco delle tabelle

Bibliografia

[1] CISIA. *Cos'è il TOLC*. https://www.cisiaonline.it/area-tematica-tolc-cisia/cose-il-tolc/. 2023.

[2] Laura Ventura e Walter Racugno. *Biostatistica. Casi di studio in R*. Egea, 2017.

[3] Alessandra Salvan, Nicola Sartori e Luigi Pace. *Modelli Lineari Generalizzati*. Springer, 2020.

[4] Richard Breen. *Regression Models: Censored, Samples Selected, or Truncated Data*. 1996.