
UNIVERSITÀ DEGLI STUDI DI MILANO

Facoltà di Scienze Politiche, Economiche e Sociali

*Master's Degree Course in Data Science for Economics*

# Statistical Analysis of the Role of the Family and School Environment in the Educational Pursuits of Adopted Children in Italy

**Advisor**: Prof. Alessandra MICHELETTI
**Co-advisor**: Prof. Silvia SALINI

**Thesis by**:
Arina LOPUKHINA
Student ID: 17169A

Academic Year 2023-2024

**Acknowledgments**

First and foremost, I would like to thank Coordinamento CARE for sharing the data and results of the surveys used by them among a set of adoptive families, on which the work of this thesis is based. Their invaluable contribution to improving adoption processes in Italy is deeply appreciated.

Furthermore, I extend my deepest gratitude to Professor Micheletti for her unwavering guidance throughout the process, as well as her patience and understanding. I also thank Professor Salini for her support during the academic journey at DSE.

Lastly, I thank my dear friends for their kindness, laughter, and companionship, which have brightened even the toughest days and made this Master's degree experience more special.

## Abstract

This study investigates the role of family and school environments as well as personal factors in shaping the educational outcomes of adopted children in Italy. Despite national policies aimed at supporting adoptees' right to education, significant gaps remain in understanding the factors influencing their academic performance. Through a mixed-methods approach combining non-parametric statistical tests, multinomial logistic regression, structural equation modeling (SEM), and advanced machine learning techniques such as Support Vector Machines (SVM) and CatBoost, this research explores key predictors of educational success and barriers faced by adoptees.

Findings suggest that age at adoption, health-related conditions, and economic independence significantly impact educational attainment. Later adoption correlates with lower educational achievement, while international adoption presents mixed effects—positively influencing secondary education but acting as a barrier to university completion. Parental education, particularly maternal education, emerges as a modest but consistent positive predictor. Furthermore, having an adopted sibling enhances resilience, particularly in accessing higher education.

While statistical and computational models provide complementary insights, challenges such as sample size limitations and class imbalances highlight the need for future research with expanded datasets and longitudinal analyses. This study underscores the importance of tailored educational policies and targeted interventions to support the unique needs of adopted students, ultimately contributing to a more inclusive and equitable educational system in Italy.

# Contents

# 1 Introduction

The research presented in this thesis extends the work of Ferritti et al. (2020), which highlights the urgent need for deeper investigation into the social integration of adoptees, particularly in relation to their educational experiences. Despite ongoing research showing the academic challenges faced by adoptees, significant gaps remain in our understanding of the complex factors influencing their school success or failure. This issue is critical because adopted children, especially those with multiple risk factors, are more susceptible to marginalization, micro-exclusions, and early school dropout—outcomes that disproportionately affect their educational and social integration.

This thesis aims to explore the various subtle, yet impactful, variables that shape the educational experiences of adoptees. These variables, including biographical fragmentation, differences from parents and peers, disabilities which impede learning, and the enduring effects of childhood adversities, require further attention to ensure equitable educational opportunities. Specifically, my research delves into how these factors influence school performance in Italy, the only country to address adoptees' right to education on a national scale through its official guidelines, Linee di indirizzo per il diritto allo studio degli alunni adottati.

Adoption represents a life-altering event for children, which causes complex psychological, emotional, and educational consequences. The integration of adoptees into society, especially within educational systems, involves a myriad of challenges that can vary based on personal, familial, and societal factors. As a data science-based exploration, this project aims to critically assess existing research on adoptee educational outcomes and the potential for data-driven analysis in understanding and improving adoptee integration. Therefore, this thesis seeks to explore the following key questions: **What are the factors within family and school contexts that most strongly predict the academic future of adopted children in Italy?** As well as **to what extent do social factors and family-school interactions collectively influence the educational achievements of adopted children in Italy?**

In order to attempt to address the above-mentioned research areas, numerous statistical methods were employed apart from descriptive analysis. Due to the complexity of the real-world survey data, it was important to concentrate on non-parametric tests and overall methods that do not require normality and homoscedasticity are tolerant

to mostly categorical variables. Furthermore, although initially believed to be a multiclass ordinal target, the violation of the proportional odds assumption indicates that educational attainment is not a straightforward ordinal process but rather one shaped by nonlinear and threshold-dependent dynamics. Hence, the influence of key predictors varies depending on the specific educational transition—whether from primary to secondary school or from secondary to higher education.

To address this, apart from Chi-Square and Kruskal–Wallis tests, the analysis incorporates Multinomial Logistic Regressions. Additionally, Structural Equation Modeling (SEM) was employed to examine the complex interrelationships between latent constructs such as physical well-being, family-school interactions, and personal features. SEM has particularly helped model indirect and mediating effects, providing deeper insights into how factors influence educational attainment. However, recognizing that parametric models may still struggle to fully account for intricate, nonlinear interactions, the study further explored robust non-parametric methods. In particular, Support Vector Machines (SVMs) and CatBoost, a gradient boosting algorithm designed for categorical data, were considered as alternative approaches. These models are particularly well-suited for analyzing high-dimensional data with complex dependencies, offering a more nuanced understanding of how social, economic, and experiential factors interact to shape educational outcomes. By integrating both traditional regression techniques and advanced computational methods, this study seeks to offer a comprehensive, data-driven analysis of the educational trajectories of adopted students, capturing both linear and non-linear dependencies in their academic experiences.

# 2 Gap of knowledge and research relevance

Adoption holds increasing social relevance in today's context, particularly in the age of declining fertility rates and shifting family dynamics. In many countries, including Italy, economic instability, precarious job markets, and the rising cost of living are leading young people to postpone starting families, contributing to demographic shifts marked by aging populations and shrinking birth rates (Organisation for Economic Co-operation and Development [OECD]). As a result, adoption is not only becoming a more prominent pathway to parenthood but also a critical area for public policy and social support systems. This evolving landscape underscores the importance of understanding how adoption affects various aspects of life, particularly in the educational sphere, where adopted children often face unique challenges that remain inadequately addressed. From a scientific perspective, there is a notable lack of quantitative social science research focused on the educational experiences of adoptees, especially those in secondary education. This gap is particularly concerning given that adolescence is a pivotal period for identity formation, including the development of adoptive identity, which is frequently overlooked in both academic literature and educational practices. My thesis aims to address this gap by providing a comprehensive analysis of the challenges faced by adoptees and adopted parents, contributing valuable data to inform more effective public policies and support systems.

A key innovation of my research lies in its methodological approach. While traditional statistical software like R and Stata are commonly used for analyzing educational outcomes, my use of Python—particularly with custom packages such as semopy for Structural Equation Modeling (SEM)—sets an interesting precedent. This choice allows for greater flexibility in data analysis, enabling more sophisticated modeling of complex relationships between variables related to emotional well-being, academic performance, and family dynamics.

Furthermore, it emphasizes the importance of cross-informant analysis, a method that remains underutilized in adoption studies. By examining data from both adoptees and their parents, my study offers a more holistic understanding of the educational challenges faced by this population. This dual perspective is crucial, as it captures the interplay between the child's personal experiences and the family environment, providing deeper insights into the factors that influence educational success.

# 3 Theoretical framework:Potential Impact of Adoption on Educational Outcomes

The adoption process in Italy is a lengthy and complex procedure, with an increasing number of adopted children—particularly those with special needs—facing significant challenges. Factors such as the origin of adoption, age at adoption, family context, learning disabilities, psychological well-being, and the school environment all play critical roles in shaping the educational outcomes of adoptees. Although Italy has implemented national guidelines, such as the Guidelines for the Right to Education for Adopted Students to support adoptees within the school system, further research is needed to assess the effectiveness of these policies. Data-driven interventions are essential to ensure better educational integration and success, especially as the number of older adoptees with special needs continues to rise. A comprehensive approach that addresses both educational and psychological needs is crucial for their long-term well-being.

Adopted children, particularly those from international adoptions, face unique hurdles in their educational development compared to their non-adopted peers. Research by Ferritti et al. (2020) highlights a correlation between adoption and lower academic performance, alongside increased behavioral challenges. These difficulties are often compounded for international adoptees, who must navigate cultural integration, cope with adverse early life experiences, and manage the long-term psychological effects of institutionalization (van IJzendoorn et. al., 2005). Howard et al. (2004) further emphasize the impact of early trauma—such as neglect and malnutrition—on cognitive development, which can significantly hinder academic progress. These issues tend to become more pronounced during adolescence and early adulthood, with Paniagua et al. (2021) noting that adoptees report higher levels of mental health concerns compared to their non-adopted counterparts.

Within the Italian context, national policies aim to mitigate these challenges by promoting the educational integration of adoptees. However, Ferritti et al. (2020) argue that more empirical evidence is needed to evaluate the real-world impact of these guidelines, especially considering the diverse backgrounds of adoptees and the multiple factors influencing their academic trajectories. Adoption in Italy is governed by a comprehensive set of rules and regulations, with a strong focus on ensuring the well-being of both children and adoptive parents. The Commissione per le Adozioni Internazionali (CAI) plays a central role in the international adoption process in

Italy, overseeing procedures and ensuring compliance with national and international guidelines. According to the latest data from 2023, Italy continues to be one of the largest receiving countries for adopted children, second only to the United States.

Nonetheless, the adoption process in Italy is characterized by a long and rigorous waiting period, which in 2023 averaged 52 months (almost four and a half years), a slight decrease compared to the previous year (CAI, 2023). This long waiting time is influenced by various factors, such as the thorough assessments of prospective adoptive parents, the complex legal procedures involved, and the matching of children with suitable families. The complexity of the process is compounded by the need to ensure that adoptive parents meet strict criteria, including financial stability, psychological readiness, and a favorable family environment. Only after fulfilling these requirements can individuals or couples be approved as adoptive parents, with regular updates and evaluations conducted throughout the adoption process.

In terms of demographics, the international adoption process in Italy shows a notable preference for boys, with males representing 58.3% of adopted children in 2023, while females accounted for 41.7% (CAI, 2023). Additionally, a significant portion of adopted children in Italy, approximately 70.4%, are classified as having special needs, which include trauma, behavioral issues, physical or mental disabilities, and learning difficulties. This shift in adoption trends, with a growing number of children with special needs, poses both challenges and opportunities for adoptee integration, particularly in the educational system.

# 4 Main Factors That Impact Adopted Children's Upbringing and Education

The educational success and overall well-being of adopted children are influenced by a variety of factors, each playing a crucial role in shaping their developmental trajectory. These factors can be broadly categorized into individual, familial, and environmental aspects, and will be further elaborated on in this section.

## 4.1 Origin of Adoption

The origin of the adopted child—whether domestic or international—has shown to have a significant impact on their upbringing and educational outcomes. International adoptees face the additional challenge of cultural integration, as they often come from different countries with diverse cultural norms, languages, and social expectations. Studies have shown that children adopted from institutions in foreign countries may experience delays in cognitive, emotional, and social development due to early neglect, malnutrition, and lack of proper care (van IJzendoorn et al., 2005). These early adversities can manifest in educational difficulties, including struggles with language acquisition, social adaptation, and emotional regulation.

In contrast, domestically adopted children may face different challenges, such as issues related to familial and psychological factors. However, they generally do not have to navigate the added complexity of cross-cultural integration. For example, Ferritti et al. (2020) emphasize that domestic adoptees may be more likely to experience challenges related to identity formation, especially when the child's background or biological family situation involves trauma or loss. Additionally, van IJzendoorn et al. (2005) found that while adopted children tend to perform better academically than those who remained in institutional care, they still face challenges in school performance compared to their nonadopted peers within the adoptive environment.

## 4.2 Age of Adoption

The age at which a child is adopted has long been recognized as one of the most critical factors influencing their educational outcomes. van IJzendoorn et al. (2005) and Cigoli and Scabini (2006) indicate that older children, particularly those adopted

after the age of seven, are more likely to have experienced early life trauma, which can manifest in behavioral problems and emotional difficulties that hinder academic performance. In Italy, an increasing number of children adopted are older than seven, with 51.7% of adoptees in 2023 falling into this age category (CAI, 2023). These children are often exposed to complex psychological factors that make their integration into the educational system more challenging.

Adopting older children also means that adoptive parents must deal with potential pre-existing academic delays and the need to provide more intensive support in terms of emotional and educational development. As these children may have had less access to early childhood education or were raised in institutional settings, they are at an elevated risk of struggling academically once they enter the school system.

## 4.3 Family Context

The family context plays a crucial role in the upbringing and educational success of adopted children. Factors such as the socio-economic status (SES) of adoptive parents, their educational background, and their ability to provide a stable and nurturing environment are central to shaping a child's educational experience. While higher SES can provide children with better access to educational resources and support systems, it does not automatically protect them from the adverse effects of early trauma (Skrondal  Laake, 2001).Welsh et al (2019) highlights the importance of parenting interventions in improving outcomes for adopted children, as targeted support can enhance parenting skills and reduce behavioral issues, although their impact on attachment security remains limited. Adoptive parents' education levels are also important, as parents with higher educational attainment are more likely to be proactive in seeking educational support for their children and providing a home environment that fosters learning. However, as Ferritti et al. (2020) note, even adoptees from well-resourced families may still face significant challenges due to the emotional and psychological impact of early adversities, which can impede their academic success.

## 4.4 Learning Disabilities

Learning disabilities are common among adopted children, particularly those who have been exposed to early adversities, such as institutionalization or neglect. Research has shown that children with a history of trauma or those who have experienced disrupted early development are more likely to exhibit learning difficulties,

including problems with language, attention, memory, and executive functioning (Cigoli and Scabini, 2006). In Italy, where a growing percentage of adopted children have special needs, including learning disabilities, the educational system faces the challenge of providing appropriate interventions to ensure these children succeed academically. Understanding the specific learning needs of adopted children and addressing them through tailored educational support is critical. This is especially true for children adopted at older ages, who may require specialized instruction to catch up with their peers in terms of basic academic skills.

## 4.5 Phychological Factors

The psychological well-being of adoptees is a key determinant of their educational success. Studies indicate that adopted children, especially those who have experienced early trauma or loss, are at a higher risk of developing psychological issues such as attachment disorders, anxiety, depression, and behavioral problems (van IJzendoorn et al., 2005; Paniagua et al., 2021). These psychological challenges can manifest in the school environment, where adoptees may struggle with self-regulation, peer relationships, and academic performance.

Adoptees' emotional adjustment often takes years, and while many children adjust well over time, others require ongoing psychological support to manage the long-term effects of their early life experiences. This highlights the need for schools to provide not only academic support but also access to mental health services that can address the specific emotional and behavioral challenges faced by adoptees. Moreover, gender has been identified as a significant factor in educational attainment among adoptees. Research suggests that female adoptees tend to perform better academically than their male counterparts, potentially due to differences in emotional regulation and social adaptation skills (van IJzendoorn et al., 2005). However, boys may be more susceptible to behavioral issues and learning difficulties, which can further hinder their academic progress.

## 4.6 School Environment

The school environment itself plays a significant role in shaping the educational outcomes of adoptees. Schools that are supportive, inclusive, and equipped to deal with the psychological and educational needs of adopted children are more likely to foster positive academic outcomes. However, as Ferritti et al. (2020) point out, there is often a lack of awareness among teachers and school staff about the specific needs

of adopted children, especially those with special needs or behavioral challenges.

In Italy, the Guidelines for the Right to Education for Adopted Students were introduced to help address these gaps. These guidelines aim to ensure that adopted children receive appropriate support in school, but more empirical research is needed to evaluate their effectiveness in practice. The social environment at school—peer relationships, teacher support, and institutional understanding of adoptee challenges—also plays a crucial role. A negative school environment, characterized by bullying, exclusion, or a lack of understanding of adoptee issues, can exacerbate psychological and academic difficulties.

# 5 Exploratory Data Analysis

This research considers two datasets, which are compilations of responses from two surveys, the first one titled "Le scelte formative degli studenti adottati" (Educational Choices of Adopted Students), is designed for parents with one or more adopted children born before 2002. Its purpose is to gather information on the educational paths of adopted children and adults to better understand their academic experiences. The questionnaire consists of 213 questions covering various topics, including demographic information, adoption details, schooling history, academic achievements, and any educational support received, such as special needs plans or certifications under Italian laws like L.104 or L.170. The second dataset is collected using the survey "Le scelte formative degli adulti adottati" (Educational Choices of Adopted Adults), which is intended for individuals who were adopted and born before 2002. Its goal was to collect information about the educational experiences of adopted individuals to better understand their academic journeys. The survey consists of 65 questions, covering demographic details, adoption history, educational background, and academic achievements. Both surveys were conducted on voluntary bases by Coordinamento CARE, with the results presented anonymously and in aggregate form.

## 5.1 Parents Survey

The initial dataset comprises 373 observations and 89 variables, capturing extensive information related to individuals' educational backgrounds, family dynamics, and socio-demographic characteristics. The variables encompass a range of data types, including integer, float, and categorical encodings, but the defeating portion of the dataset is categorical variables, often ordinal or binary. They include identifiers such as critical demographic information like gender, year of birth, and adoption status, detailed educational metrics, including school diploma scores, as well as higher education achievements through variables like bachelor's/master's degree attainment. Family background information is robustly represented, with variables detailing parental education levels, parental involvement in school activities, and familial support structures. The dataset also provides insights into special educational needs and challenges, such as individualized education plans for special needs presence of certifications for learning disabilities, such as dyslexia or ADHD. Predictors

were assessed and selected based on data quality and availability. Unfortunately, despite some questions being extremely interesting to explore, sometimes there were too many missing variables or the inputs were too complex to encode while preserving the meaning and not creating multicollinearity issues. Also, given the limited sample size, the number of features needed to be limited to minimize the data noise and ensure algorithm convergence. Hence, the final version of the dataset only contained 21 variables, which was further reduced for some of the models' design. Exploring contingency tables, some initial patterns emerged. For instance, while some regions of origin exhibit a relatively balanced distribution between men and women, others reveal notable disparities that may reflect historical adoption trends, regional adoption policies, or socio-cultural preferences in adoptive families. For adoptees from Africa and Asia, the gender distribution is nearly equal, with 10 men and 11 women from Africa and 31 men and 32 women from Asia. This suggests that adoptions from these regions were not significantly influenced by gender preference and that adoption rates remained fairly balanced for both boys and girls.

A striking gender disparity is observed in Eastern Europe, where 83 men and only 35 women were adopted. This pattern may be influenced by country-specific adoption policies, cultural attitudes toward gender, or even the availability of children for adoption in these regions at different points in time. Historically, some Eastern European countries experienced surges in international adoptions due to socio-political instability, which may have disproportionately affected boys due to institutional biases or demographic factors within orphanages. In contrast, adoptees from Italy present the opposite trend, with more women (47) than men (31). Given that domestic adoptions in Italy are more regulated and involve fewer socio-political barriers than international adoptions, the gender imbalance has likely occured by chance. For adoptees from Latin America, the trend shifts again, with 53 men and 33 women. While the disparity is not as extreme as in Eastern Europe, it still suggests a higher adoption rate of boys compared to girls. Overall, these findings suggest that gender disparities in adoption vary significantly by region, likely shaped by historical, social, and policy-driven factors. The overrepresentation of men in adoptions from Eastern Europe and Latin America contrasts with the more balanced distributions in Africa and Asia, while Italy stands out for having more female adoptees.

Furthermore, the data reveals distinct patterns in educational attainment by gender, highlighting a shift in academic progression between men and women. At lower levels of education, men are more represented than women. For instance, more men (22) than women (12) have only completed middle school, and similarly, 33 men hold a non-university professional diploma compared to just 17 women.

Table 5.1: Gender Distribution by Region of Origin

| **Origin Region** | **Male (0)** | **Female (1)** |
|---|---|---|
| Africa | 10 | 11 |
| Asia | 31 | 32 |
| East Europe | 83 | 35 |
| Italy | 31 | 47 |
| Latin America | 53 | 33 |
| Other | 4 | 3 |

This suggests that men are more likely to pursue vocational education or enter the workforce earlier rather than continuing into higher academic qualifications. At the high school level, both genders reach their highest numbers, with 132 men and 94 women obtaining a diploma that grants access to university. This indicates that secondary education is the most commonly completed level for both groups. However, beyond this stage, a noticeable shift occurs, as women begin to surpass men in higher education attainment. While 16 men have a university degree, this number rises to 27 among women. The same trend extends to postgraduate education, where six women hold a post-university qualification compared to only three men.

Table 5.2: Educational Attainment by Gender

| **Educational Attainment** | **Male (0)** | **Female (1)** |
|---|---|---|
| No title | 0 | 1 |
| Middle school | 22 | 12 |
| Professional diploma (non-university) | 33 | 17 |
| High school diploma | 132 | 94 |
| IFTS | 6 | 4 |
| University degree | 16 | 27 |
| Advanced university degrees (Master's) | 3 | 6 |

These findings suggest a gendered divergence in educational paths. While men are more likely to stop at lower education levels or opt for vocational pathways, women are increasingly continuing into higher education, including university and postgraduate studies. This pattern aligns with broader global trends, where female participation in tertiary education has risen significantly over recent decades. However, the fact that men are overrepresented in lower education categories raises concerns about early disengagement from academic progression, suggesting a need for further investigation into the social or economic factors influencing these decisions.

On the other hand, with respect to parental level of education versus the one of

the adopted children, among fathers and mothers, there is a significant peak at the high school diploma level, followed by a peak at the university degree level. The adoptees population follows the same structure but with a more pronounced peak at the middle school and high school diploma levels, suggesting a slightly lower overall educational attainment. The following figures illustrate the distribution of educational attainment for mothers, fathers, and the adopted children, with the levels of education encoded as can be seen below:

- 0 = No title/Elementary school certificate or attendance at only primary school
- 1 = Secondary school diploma (previously known as lower secondary school certificate)
- 2 = Professional qualification diploma that does not allow access to university (3 or 4 years)
- 3 = Upper secondary school diploma (previously known as high school diploma) that allows university enrollment (5 years)
- 4 = Post-secondary non-tertiary qualification (IFTS)
- 5 = University degree
- 6 = Postgraduate degree (Master's, Advanced training school, etc.)
- 7 = Doctoral degree (PhD)

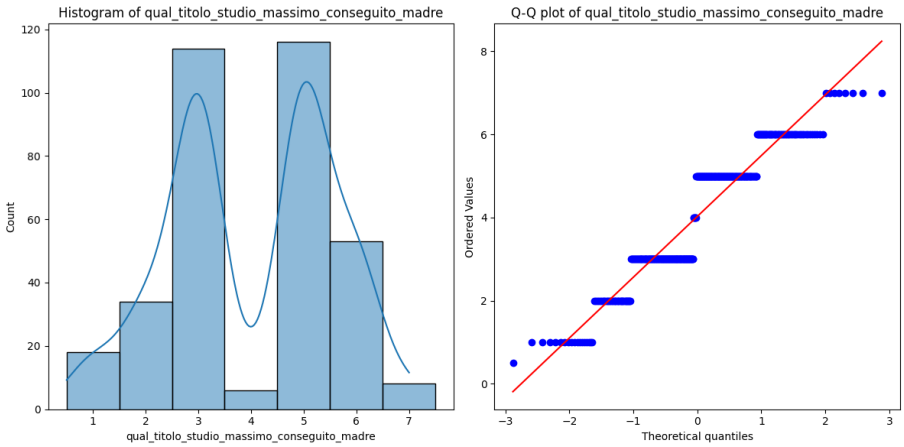



Figure 5.1: Histogram and Q-Q plot of maternal educational attainment.

The maternal education histogram (Figure 5.1) shows a multimodal distribution with significant peaks at high school diploma and Bachelor's degree levels. The Q-Q plot confirms non-normality, indicating categorical clustering.

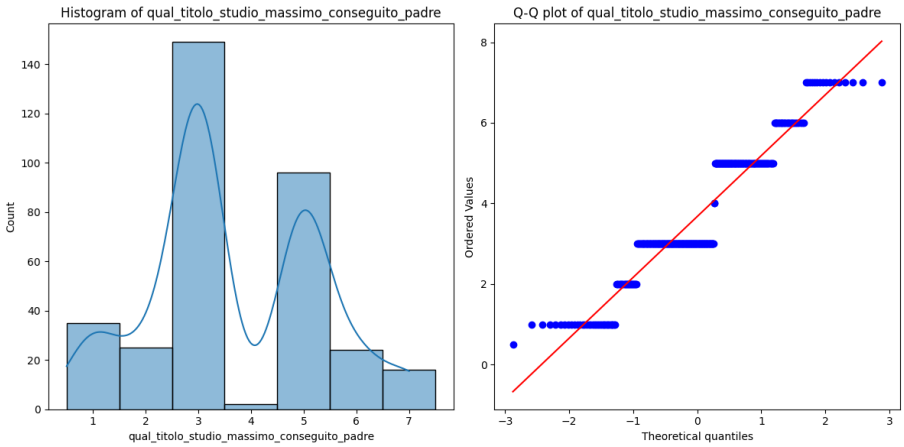



Figure 5.2: Histogram and Q-Q plot of paternal educational attainment.

As shown in Figure 5.2, fathers' educational attainment follows a similar trend, with a strong peak at high school diplomas. Although there are less Master's graduates than among mothers, a slightly higher percentage hold advanced PhD degrees compared to the latter.

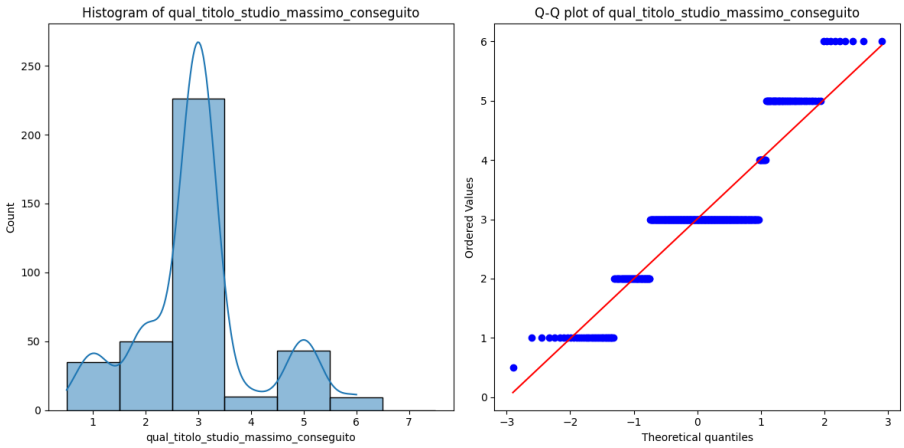



Figure 5.3: Histogram and Q-Q plot of educational attainment among adoptees.

Figure 5.3 summarizes the educational distribution among adoptees, indicating the contrast compared trends observed in parental education, as most of the observations are concentrated in the level corresponding to High School Diploma, yet less than 50 adoptees have some type of a university degree.

Lastly, to assess the initial correlations before applying any statistical tests, Pearson and Spearman correlations matrices were examined. Although no unexpected patterns could be immediately identified, the Spearman's correlation matrix highlights key relationships among socio-educational variables. **Economic independence** strongly correlates with **employment status** ($p = 0.64$), while **school performance**

**(final grade score)** is negatively linked to **employment** (p = -0.26), suggesting work may impact academic outcomes. **Gender differences** are minimal, though females show a slight tendency toward higher educational attainment. **International adoption** negatively correlates with **first school grade attended** (p = -0.32), reflecting differences in educational integration.

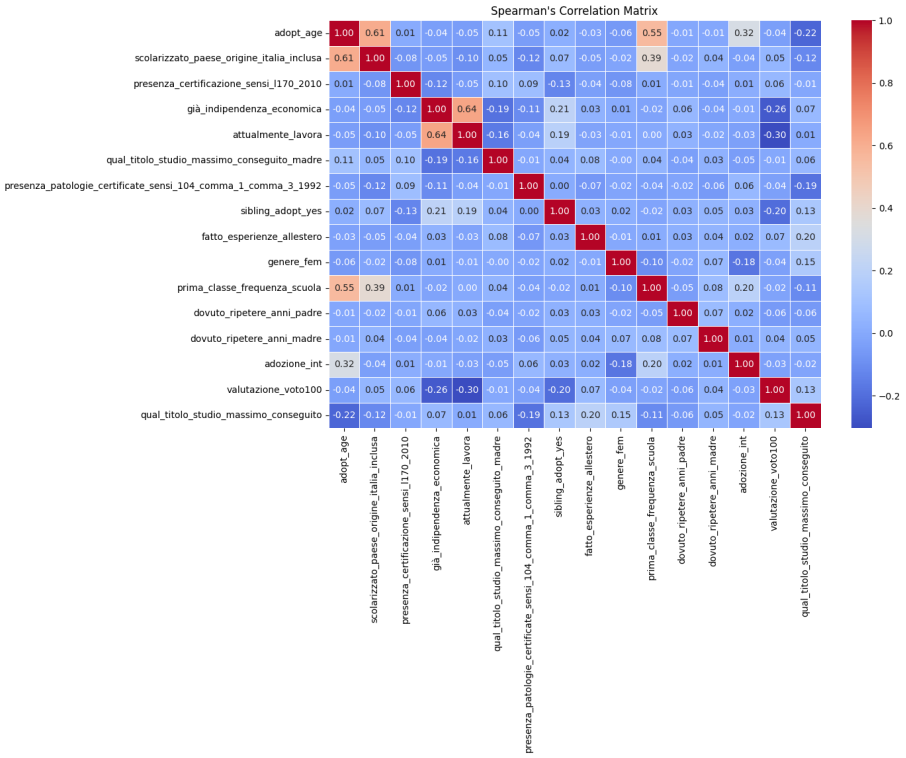



Figure 5.4: Spearman's Correlation Matrix - Parents

Overall, the findings emphasize the necessity of using statistical models that account for the ordinal nature of educational attainment, while also highlighting the broader societal trends that shape educational achievement. Nonetheless, more substantial conclusions to be derived from application of statistical modeling, as EDA cannot be conclusive.

## 5.2 Adopted Children Survey

With respect to the second dataset, it is largely similar in structure to the parents' one as the survey questions were mirrored. It comprises 105 entries with 66 variables with similar demographic, educational, and socioeconomic characteristics of individuals. Like in case of the parents' dataset, the data spans various domains such as personal background (e.g., year of birth, country of origin, year of arrival in the family), educational trajectory (e.g., age at school entry, highest education

level achieved, university experiences), and family context (e.g., parents' educational background and professions). Key features include both quantitative data (e.g., ages, high school grades out of 100, satisfaction scores) and categorical data (e.g. gender, adoption type, educational qualifications). The dataset also contains variables related to learning difficulty, health certifications, and economic independence, just like in case of the first dataset.

Notably, this dataset contains Likert scale answers regarding whom the adoptees consider influential figures on their educational paths. Thus, we can see to what extent teachers, parents, or friends played an influential role in their school journey. Another meaningful piece of information is the children's self-reported perception of whether their experience at school as well as university was harder or easier compared to their peers.

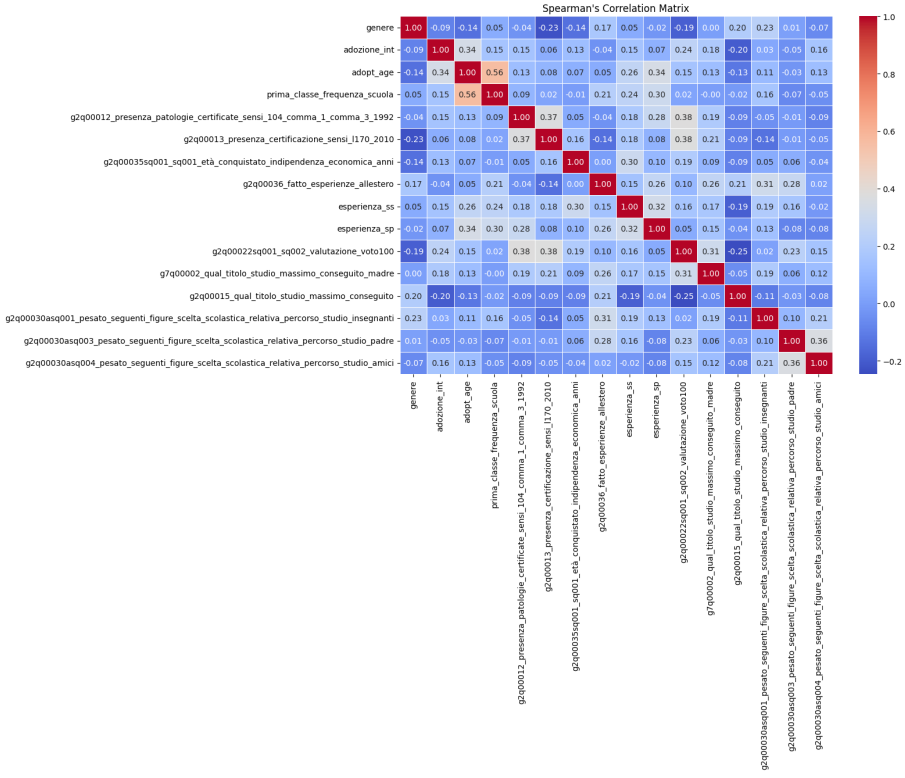



Figure 5.5: Spearman's Correlation Matrix - Children

The Spearman correlation matrix provides insight into the relationships between various predictors, highlighting both direct associations and potential indirect influences. Similarly as for the other dataset, the correlation between adoption age and educational attainment is negative, suggesting that individuals adopted at an older age tend to achieve lower levels of education. Also, economic independence exhibits a negative correlation with education, reinforcing the idea that early financial

independence may come at the expense of prolonged schooling. Parental education, particularly the mother's highest level of education, shows a weak but positive correlation with educational attainment, indicating that while parental education does exert some influence, it is not the dominant factor in determining a child's academic progression.

Moreover, studying or working abroad exhibits a weak but positive association with educational attainment, aligning with previous findings that international exposure can contribute to higher academic success. Interestingly, gender does not exhibit a strong correlation with education, suggesting that any potential disparities in academic outcomes between male and female students are either minimal or driven by other confounding factors. Learning disabilities, measured by the presence of certified pathologies and specific learning disorders, do not show a strong association with education levels.

Lastly, the correlation between the perceived influence of teachers and educational attainment is negative, suggesting that a greater perceived influence of teachers on educational choices might be associated with lower levels of attainment, potentially due to external pressures or a lack of intrinsic motivation. In contrast, the influence of parents shows a weak positive correlation, implying that parental involvement may provide a modest boost to educational achievement.

# 6 Methodology

This section describes the applied algorithms and the metrics used for their evaluation. This study employs a comprehensive statistical analysis framework to investigate the educational outcomes of adopted children, while recognizing the dataset's complexities—including a relatively small sample size, unbalanced class distribution, predominantly categorical data, non-normal distributions, heteroscedasticity, and the presence of outliers—the methodological approach has been meticulously designed to address these challenges and ensure robust, valid results.

## 6.1 Prevailing Analytical Approaches in Adoption Research

Historically, research on the educational outcomes of adopted children has predominantly utilized descriptive statistics and basic regression analyses. For instance, studies have mainly employed linear regression models to explore the relationship between adoptees' academic achievements and parental education levels (Plug  Vijverberg, 2003). Similarly, analyses of national survey data have often relied on straightforward statistical comparisons to highlight disparities in school performance between adopted and non-adopted children (Tartaglia  Rankin, 2023). While these methods have been instrumental in identifying general trends, they may not fully capture the intricate interplay of factors influencing adoptees' educational trajectories.

## 6.2 Statistical Techniques and Model Specification

To build upon and enhance the existing body of research, this study integrates advanced analytical techniques—Structural Equation Modeling (SEM), Support Vector Machines (SVM), and CatBoost—that offer deeper insights into the multifaceted determinants of educational outcomes among adopted children.

### 6.2.1 Chi-Square and Kruskal-Wallis Tests

For categorical and ordinal data, we first employ chi-square tests and Kruskal-Wallis tests to assess relationships between variables. The chi-square test ($\chi^2$) evaluates whether observed frequencies significantly differ from expected frequencies under the null hypothesis of independence of variables:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{6.1}$$

where $O_i$ represents observed frequencies, and $E_i$ represents expected frequencies. The Kruskal-Wallis test is applied as a non-parametric alternative to ANOVA:

$$H = \frac{12}{N(N+1)} \sum R_j^2 - 3(N+1) \tag{6.2}$$

where $R_j$ is the sum of ranks for group $j$, and $N$ is the total number of observations.

### 6.2.2 Ordered Logistic Regression (L1 and L2-Regularized)

To model the relationship between educational attainment (ordinal dependent variable) and multiple predictors, we apply ordered logistic regression with L1 and L2 regularization. The ordered logit model estimates the probability that an individual falls into category $k$ or below:

$$P(Y \leq k) = \frac{1}{1 + \exp(-(\alpha_k - \beta X))} \tag{6.3}$$

where $\alpha_k$ are threshold values for categories, and $\beta X$ represents the predictor variables. Regularization is introduced to mitigate overfitting:

$$L(\beta) = -\sum [y_i \log P_i + (1 - y_i) \log(1 - P_i)] + \lambda \sum \beta_j^2 + \alpha \sum |\beta_j| \tag{6.4}$$

where $\lambda$ is the L2 regularization parameter and $\alpha$ is the L1 regularization parameter. The Ordered Logistic Regression model is implemented using the *statsmodels* package, specifically the OrderedModel class, while L1 and L2 regularization are applied using the *sklearn.LogisticRegression* with solver='saga' to allow for both types of penalties.

### 6.2.3 Multivariate Logistic Regression

For categorical outcomes, a multivariate logistic regression model is specified as follows:

$$P(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)} \tag{6.5}$$

where $\beta_0$ is the intercept and $\beta_p$ represents the coefficient estimates for predictor variables. The model was implemented using the MNLogit class from the *statsmodels* package in Python.

To assess the performance of the regression models, several evaluation metrics were employed. Pseudo R-squared (McFadden's $R^2$) was used to measure goodness-of-fit by comparing the log-likelihood of the fitted model to that of a null model, providing insight into how well the independent variables explain the variation in educational attainment. Additionally, Mean Absolute Error (MAE) was calculated to quantify the average deviation of predicted education levels from actual values, helping to assess the practical accuracy of the model. Furthermore, a confusion matrix with adjacent accuracy was analyzed to determine the proportion of cases where predictions were within one category of the true education level, recognizing that small classification errors in an ordinal setting may still provide useful information about a participant's educational trajectory. These metrics together aim to provide a comprehensive evaluation of the model's predictive power and practical utility.

### 6.2.4 Structural Equation Modeling (SEM)

From a data science perspective, addressing the complexity of adoptees' educational outcomes requires advanced, multifaceted analytical approaches. Statistical methods such as Structural Equation Modeling (SEM) and factor score analysis can help account for latent variables like emotional well-being, early adversity, and academic performance, providing a comprehensive understanding of how adoption affects educational integration. SEM helped facilitate the examination of both direct and indirect effects while integrating latent constructs and accounting for measurement error. This approach is particularly advantageous in modeling complex relationships in social science research, such as the interplay between educational attainment, psychological well-being, and socio-economic factors. By quantifying mediating effects, SEM elucidates pathways through which parental support and early developmental experiences influence adoptees' academic performance (Mîndrilă, 2010). The general form is:

$$Y = \Lambda\xi + \delta, \quad X = \Gamma\eta + \zeta \tag{6.6}$$

where $Y$ represents observed endogenous variables, $X$ represents exogenous variables, $\Lambda$ and $\Gamma$ are coefficient matrices, and $\delta$ and $\zeta$ denote residual terms. The models were implemented thanks to the *semopy* Python package and model fit was

later assessed using indices such as RMSEA, chi-squared p value, CFI, and TLI.

### 6.2.5 Support Vector Machines (SVM)

In addition to traditional statistical methods, machine learning techniques are introduced to enhance predictive accuracy and classification performance. The use of Support Vector Machines ensures that complex, non-linear relationships between predictors and educational outcomes are adequately captured. SVMs are employed to improve classification tasks, especially given the challenges posed by high-dimensional categorical data, as they have been effectively applied in predictive modeling tasks involving small and unbalanced datasets, making them appropriate for this study (Kong et al., 2020). For the tasks outlined in this thesis, SVM was employed using a radial basis function (RBF) kernel, implemented using the SVC class from the *sklearn.svm* module in Python:

$$K(x_i, x_j) = \exp(-\gamma |x_i - x_j|^2) \tag{6.7}$$

where $\gamma$ is the kernel parameter. The SVM objective is to find the optimal hyperplane $w^T x + b = 0$ that maximizes the margin while minimizing the loss function:

$$\min \frac{1}{2}|w|^2 + C \sum \xi_i \tag{6.8}$$

where $C$ is the regularization parameter and $\xi_i$ are slack variables for misclassification.

### 6.2.6 CatBoost Algorithm

To further enhance model interpretability and predictive power, the CatBoost algorithm is utilized. CatBoost, a gradient boosting method specifically designed for categorical data, processes such variables without requiring extensive preprocessing. Unlike traditional boosting methods, CatBoost effectively handles categorical features without excessive reliance on one-hot encoding, reducing dimensionality and mitigating overfitting issues. Research on CatBoost has demonstrated its superior performance in structured data applications, particularly in handling data with heterogeneous categorical features (Ostroumova et al., 2017). Integrating CatBoost into the analytical pipeline was done using the *CatBoostClassifier* class from the

*catboost* package. The loss function optimized in CatBoost is the log-loss function for classification:

$$L = -\sum [y_i \log P(y_i) + (1 - y_i) \log(1 - P(y_i))] \tag{6.9}$$

This method minimizes overfitting through ordered boosting and uses oblivious trees for structured feature selection (Ostroumova et al., 2017).

Overall, the study acknowledges and addresses several methodological challenges:

**1. Small Sample Size and Unbalanced Class Distribution**: The use of regularization techniques in regression models and the application of machine learning algorithms like SVM and CatBoost help mitigate issues arising from small sample sizes and unbalanced class distributions.

**2. Predominantly Categorical Data**: Methods such as chi-square tests, Kruskal-Wallis tests, and the CatBoost algorithm are specifically chosen for their proficiency in handling categorical data effectively.

**3. Non-Normal Distributions and Heteroscedasticity**: The adoption of non-parametric tests and robust modeling techniques, including SEM, ensures that the analyses remain valid despite deviations from normality and homoscedasticity assumptions.

**4. Outliers:** Sensitivity analyses are conducted to assess the impact of outliers, and robust statistical methods are employed to minimize their influence on the results.

By integrating both traditional statistical techniques and advanced machine learning methods, this study provides a nuanced analysis of the factors influencing educational outcomes among adopted children. This multifaceted approach ensures a rigorous examination of the data, aligning with best practices in quantitative social science research, and contributes valuable insights to the existing literature on adoption and education.

# 7 Data Preprocessing

After selecting theoretically-relevant predictor variables, in order to ensure data quality and enhance model performance, several preprocessing techniques were implemented to address issues related to categorical encoding, dimensionality reduction, class imbalance, missing values, and standardization. Given the complexity of the dataset, which included a mixture of ordinal and nominal categorical variables, a tailored encoding approach was necessary. Ordinal variables were transformed using ordinal encoding to preserve inherent category order, while nominal variables were processed with frequency encoding to prevent the introduction of artificial ordinal relationships.

To reduce data sparsity and improve interpretability, dimensionality reduction techniques were applied where feasible. One of the most significant modifications involved consolidating detailed country names into broader geographical categories such as continents. This transformation preserved essential information while reducing unnecessary complexity, ultimately enhancing the efficiency of the model. Also, feature engineering was employed to extract more meaningful insights from existing variables. One particularly relevant feature introduced was the age at adoption, calculated as the difference between the year of birth and the year of adoption. This variable was included to assess whether the timing of adoption influences long-term educational outcomes, particularly given existing research suggesting that late adoption can be associated with greater academic challenges.

Firstly, both the adult and children survey datasets present several methodological challenges that must be carefully considered during analysis. One of the most prominent issues is the high proportion of missing data, with numerous NA values scattered across various variables. This missingness is supposedly due to participants skipping or selectively ignoring certain questions, which may introduce bias if the missing data is not random. To address the issue of missing data in both the adult and children survey datasets, several preprocessing steps were undertaken. Instead of removing the rows entirely, a placeholder value of mean or mode to represent missing responses was used. This method preserved the already scarce dataset structure while allowing the missingness to be explicitly accounted for during the analysis.

Another critical concern is the imbalance between the number of variables and the sample size. Both datasets contain an extensive range of variables relative to the limited number of observations, increasing the risk of overfitting in statistical models

and reducing the reliability of inferential conclusions. This high dimensionality problem necessitates strategic reduction in the number of variables through feature selection methods or dimensionality reduction techniques, focusing only on the most relevant columns to maintain statistical power. For instance, the majority of observations in target variable fell into class 3, representing mid-level secondary education, while higher education categories, particularly university degrees (classes 5, 6, and 7), were substantially underrepresented. Without correction, this imbalance would have led to biased models that favored the majority class while failing to learn meaningful patterns from the minority categories. To mitigate this, *SMOTETomek* was applied, leveraging a combination of Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for underrepresented classes and Tomek links to remove overlapping instances from the majority class. This dual approach not only balanced class distributions but also eliminated noisy observations that could hinder classification performance. However, in both datasets, categories representing higher education (bachelor's, master's, PhD) had to be combined into one class, as upsampling was not yielding accurate results. Lastly, it is important to note that upsampling was only applied to the training sets as upsampling the entire population could lead to data leakages and hence introduce bias to the models. This created accuracy constraints since some classes were underrepresented in the test samples, leading to lower prediction precision.

Moreover, the data in both datasets deviates from normal distribution assumptions. Many variables exhibit non zero skewness and kurtosis different from the standard normal one , which can violate the assumptions of parametric tests commonly used in statistical analyses, such as linear regression or ANOVA. To address these issues, StandardScaler was applied to normalize continuous predictors. By transforming each variable to have a mean of zero and a standard deviation of one, this approach ensured that features with larger magnitudes did not dominate model performance. Unlike alternative scaling methods, such as min-max scaling, standardization preserved the relative distribution of the data, making it particularly suitable for models that assume normally distributed inputs.

Furthermore, heteroscedasticity is evident in both datasets, as the variance of the residuals is not constant across different levels of the independent variables. As violation of homoscedasticity assumptions can result in inefficient estimators and biased standard errors, compromising the validity of hypothesis tests, to account for this issue, it was necessary to opt for robust estimators that adjust for heteroscedasticity while employing parametric methods, ensuring more reliable coefficient estimates and accurate inference.

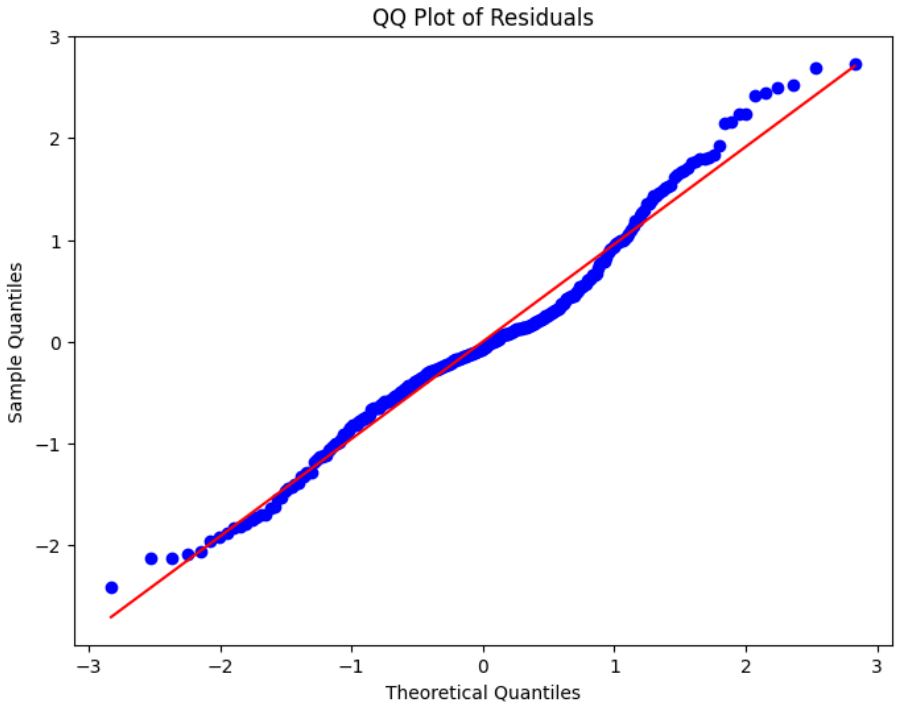



Figure 7.1: QQ Plot - Parents' Dataset

Lastly, there were signs of multicollinearity that were aggravated due to small sample size, even if in some cases excessive correlation did not make sense from a theoretical standpoint. Running variance inflation factor (VIF) tests on both datasets helped to identify the problematic sets of variables, such as, for instance, multicollinear relationship between variables related to mothers and father's responses. This led to elimination of some variables, which initially seemed significant from the theoretical standpoint, in order to ensure algorithm convergence and interpretability of model results.

# 8 Results

In this section, the performances of the mentioned methods in the previous chapter are presented and analyzed.

## 8.1 Preliminary assessment

### 8.1.1 Chi-Squared Test

The chi-square test for independence was conducted separately for adult and children's datasets to examine the statistical relationships between various categorical variables and educational attainment. Unlike tests that assume normality, the Chi-Square test does not require assumptions about the distribution of the variables involved, which makes it particularly useful in case of our data. For both datasets, the target variable was identified as the level of education, with the following categories:

- No title/Elementary school certificate or attendance at only primary school
- Secondary school diploma (previously known as lower secondary school certificate)
- Professional qualification diploma that does not allow access to university (3 or 4 years)
- Upper secondary school diploma (previously known as high school diploma) that allows university enrollment (5 years)
- Post-secondary non-tertiary qualification (IFTS)
- University degree
- Postgraduate degree (Master's, Advanced training school, etc.) – combined with Doctoral degree (PhD)

One of the key areas of investigation was the role of parental education in shaping academic success. In the adult dataset, the relationship between an individual's highest level of education and their mother's education yielded a chi-square statistic of 28 with a p-value of 0.46, indicating no statistically significant relationship. Similarly, in the children's dataset, parental education showed non-significant p-values of 0.21 and 0.15, suggesting that, at least within these datasets, there is

Table 8.1: Chi-Square Test for Independence Results - Parents' Dataset

| Variables | Statistic | p-value | Degrees of Freedom |
|---|---|---|---|
| qual_titolo_studio vs scolarizzato_paese_origine | 6.98 | 0.137 | 4 |
| qual_titolo_studio vs presenza_certificazione | 5.31 | 0.257 | 4 |
| qual_titolo_studio vs indipendenza_economica | 35.93 | $2.99 \times 10^{-7}$ | 4 |
| qual_titolo_studio vs attualmente_lavora | 35.67 | $3.39 \times 10^{-7}$ | 4 |
| qual_titolo_studio vs titolo_studio_madre | 27.95 | 0.467 | 28 |
| qual_titolo_studio vs patologie_certificate | 16.25 | 0.0027 | 4 |
| qual_titolo_studio vs sibling_adopt_yes | 23.27 | $1.12 \times 10^{-4}$ | 4 |
| qual_titolo_studio vs esperienze_allestero | 16.62 | 0.0023 | 4 |
| qual_titolo_studio vs genere_fem | 11.19 | 0.0245 | 4 |
| qual_titolo_studio vs ripetere_anni_padre | 3.04 | 0.551 | 4 |
| qual_titolo_studio vs ripetere_anni_madre | 8.64 | 0.374 | 8 |
| qual_titolo_studio vs adozione_int | 1.75 | 0.782 | 4 |

Table 8.2: Chi-Square Test for Independence Results - Children's Dataset

| Variables | Statistic | p-value | Degrees of Freedom |
|---|---|---|---|
| g2q00015_qual_titolo_studio vs genere | 6.64 | 0.0844 | 3 |
| g2q00015_qual_titolo_studio vs adozione_int | 6.35 | 0.0960 | 3 |
| g2q00015_qual_titolo_studio vs prima_classe_frequenza_scuola | 12.51 | 0.820 | 18 |
| g2q00015_qual_titolo_studio vs presenza_patologie | 7.57 | 0.0558 | 3 |
| g2q00015_qual_titolo_studio vs presenza_certificazione | 7.57 | 0.0558 | 3 |
| g2q00015_qual_titolo_studio vs esperienze_allestero | 14.90 | 0.0019 | 3 |
| g2q00015_qual_titolo_studio vs esperienza_ss | 8.60 | 0.1975 | 6 |
| g2q00015_qual_titolo_studio vs esperienza_sp | 3.45 | 0.7501 | 6 |
| g2q00015_qual_titolo_studio vs qual_titolo_studio_madre | 22.30 | 0.2190 | 18 |
| g2q00015_qual_titolo_studio vs scelta_studio_insegnanti | 13.61 | 0.1370 | 9 |
| g2q00015_qual_titolo_studio vs scelta_studio_padre | 10.63 | 0.3023 | 9 |
| g2q00015_qual_titolo_studio vs scelta_studio_amici | 6.19 | 0.7212 | 9 |

no strong direct statistical association between parental education and a child's educational outcomes. However, this does not imply that parental education has no influence; rather, its impact may be more indirect, potentially shaping academic performance through socioeconomic status, access to educational resources, and parental support. Interestingly, the presence of an adopted sibling was shown to have statistical significance, as, perhaps, having a close relative going to a similar educational experience can provide more support and help a child navigate their educational path.

Another key factor analyzed was gender and its role in educational attainment, with mixed results across the two datasets. In the adult dataset, gender showed a significant relationship ($p = 0.024$) with education level, suggesting that educational trajectories differ by gender among adults. In contrast, in the children's dataset, gender did not exhibit a highly significant association with academic attainment ($p = 0.084$). Similarly, international adoption status was tested in both datasets, and in both cases, it showed no significant effect on educational outcomes ($p = 0.78$ in

adults, p = 0.096 in children). This suggests that being adopted from abroad does not, in itself, serve as a determining factor in a student's academic success.

A notable difference between the two datasets emerged in the relationship between learning disabilities and educational attainment. In the adult dataset, individuals with medically certified conditions under Law 104/1992 showed a significant association with lower levels of education, indicating that learning disabilities may present substantial academic challenges that persist into adulthood. In contrast, the children's dataset showed a weaker but still notable relationship (p = 0.055) between learning disabilities (defined under Law 104/1992 and Law 170/2010 certifications) and academic performance.

One of the strongest and most consistent findings in both datasets was the significant relationship between experiential learning opportunities and educational achievement. In both the adult and kids datasets, individuals who had studied or lived abroad for extended periods were more likely to attain higher education levels (p = 0.0022 in adults, p = 0.0019 in children). This suggests that international exposure fosters academic success, potentially due to broader cultural perspectives, language acquisition, and enriched learning experiences. However, self-assessed learning struggles among adopted children did not show a statistically significant effect on their academic performance. This indicates that while some adopted students may perceive themselves as having faced greater academic challenges compared to their peers, these perceived struggles do not necessarily translate into significantly lower educational outcomes.

Overall, while economic independence and employment are highly predictive of education level in the parents' dataset, children's academic outcomes appear to be more influenced by learning opportunities and, to a lesser extent, learning disabilities. Interestingly, while parental education does not show a direct statistical correlation with children's academic performance, its influence may still be embedded in other factors such as financial stability and parental support systems.

### 8.1.2 Kruskal-Wallis Test

Additionally, the Kruskal-Wallis test has been performed as a non-parametric alternative used to compare distributions of a categorical variable across multiple groups, once again, without assuming that the data follows a normal distribution. The test is a non-parametric alternative to the ANOVA test, used to compare three or more independent groups when the assumption of normality is not met. It offers a different perspective from the chi-square analysis, allowing us to assess whether differences in educational attainment exist between groups rather than just testing

for independence.

One of the similarities between the results of the two tests is the lack of a direct statistical relationship between parental education and children's educational attainment. In the chi-square test, neither mother's nor father's education showed significant associations with the child's highest level of schooling, and the Kruskal-Wallis test confirmed this, with p-values of 0.48 (parents) and 0.33/0.19 (children). The results for gender also align closely between the two tests, as the Kruskal-Wallis test found that gender differences in education were statistically significant for both generations, though weaker among children ($p = 0.043$). This suggests that while gender disparities in education are diminishing in the younger generation, they are still present and may become more pronounced as these students progress into higher education or the workforce.

The impact of foreign adoption status was another area where the two tests produced slightly different results. In the chi-square test, type of adoption did not have a significant association with education levels in either dataset, suggesting that being adopted from another country did not systematically influence academic outcomes. However, in the Kruskal-Wallis test, adoption status was found to be significant in the children's dataset ($p = 0.039$), indicating that there are measurable differences in education levels between adopted and non-adopted children. This suggests that while adoption status may not strongly predict whether a child will succeed academically, adopted students may experience different educational trajectories that lead to variations in attainment. Similarly, the influence of learning disabilities on education showed some agreement between the two tests. Both the chi-square and Kruskal-Wallis results indicated that learning disabilities under Law 104/1992 were significantly associated with lower educational attainment. Notably, disabilities such as dyslexia or dysgraphia were not found significant in either of the datasets.

One of the most consistent findings across both tests was the impact of experiential learning opportunities, particularly studying abroad. In both the chi-square and Kruskal-Wallis analyses, having studied or worked abroad was strongly associated with higher education levels in both the parents' and children's datasets ($p = 0.0022$ and $0.0019$ in chi-square; $p = 0.0001$ and $0.034$ in Kruskal-Wallis). This suggests that international exposure plays a crucial role in academic success across generations, possibly due to cultural immersion, language skills, and expanded learning opportunities. The consistency of this finding across both statistical tests strongly supports the idea that students who gain global educational experiences are more likely to have higher levels of schooling.

Table 8.3: Kruskal-Wallis Test Results - Parents' Dataset

| **Variable** | **Statistic** | **p-value** |
|---|---|---|
| qual_titolo_studio vs scolarizzato_paese_origine | 5.01 | 0.0252 |
| qual_titolo_studio vs presenza_certificazione | 0.04 | 0.8361 |
| qual_titolo_studio vs indipendenza_economica | 1.67 | 0.1964 |
| qual_titolo_studio vs attualmente_lavora | 0.01 | 0.9165 |
| qual_titolo_studio vs titolo_studio_madre | 6.50 | 0.4830 |
| qual_titolo_studio vs patologie_certificate | 13.07 | 0.0003 |
| qual_titolo_studio vs sibling_adopt_yes | 6.58 | 0.0103 |
| qual_titolo_studio vs esperienze_allestero | 14.52 | 0.0001 |
| qual_titolo_studio vs genere_fem | 8.21 | 0.0042 |
| qual_titolo_studio vs ripetere_anni_padre | 1.41 | 0.2350 |
| qual_titolo_studio vs ripetere_anni_madre | 0.76 | 0.6850 |
| qual_titolo_studio vs adozione_int | 0.20 | 0.6577 |

Table 8.4: Kruskal-Wallis Test Results for Educational Attainment - Children's dataset

| **Variable** | **Statistic** | **p-value** |
|---|---|---|
| genere | 4.077 | 0.043 |
| adozione_int | 4.227 | 0.039 |
| prima_classe_frequenza_scuola | 3.286 | 0.772 |
| presenza_patologie_certificate | 0.901 | 0.343 |
| presenza_certificazione_sensi_l170 | 0.901 | 0.343 |
| fatto_esperienze_allestero | 4.453 | 0.035 |
| esperienza_ss | 5.682 | 0.058 |
| esperienza_sp | 1.771 | 0.412 |
| titolo_studio_madre | 6.853 | 0.335 |
| scelta_scolastica_insegnanti | 6.076 | 0.108 |
| scelta_scolastica_padre | 0.351 | 0.950 |
| scelta_scolastica_amici | 2.838 | 0.417 |

While the chi-square and Kruskal-Wallis tests largely reinforce each other, one of the main differences is that the Kruskal-Wallis test is more sensitive to differences between groups rather than just detecting whether two variables are associated. This explains why adoption status was significant in the Kruskal-Wallis test for children but not in the chi-square test, as it indicates that adopted children do have different education levels even if the variable itself is not completely independent of educational attainment. Similarly, the chi-square test suggested that economic independence and employment were highly predictive of education level in the adult dataset ($p < 0.001$), but in the Kruskal-Wallis test, these same variables did not show significant differences ($p = 0.19$ for economic independence, $p = 0.91$ for employment status).

This suggests that while there may be a strong relationship between education and employment, education levels themselves may not vary as drastically across different employment or economic status groups, which calls to explore multinomial methods.

## 8.2 Logit Models

While Chi-Squared and Kruskal-Wallis tests were interesting starting points for identifying high-level relationships between dependent and independent variables, logistic regression models were employed as it did not require the assumption of normally distributed residuals and was better suited for handling categorical and ordinal outcome variables. Also, L1/L2 regularization and robust standard errors (HC3) were applied to account for heteroscedasticity, and interaction terms were explored to capture potential non-linear relationships. While these adjustments improved the reliability of the model, the limitations imposed by the data structure remained a key consideration throughout the analysis, underscoring the complexity of studying educational outcomes among adopted students using real-world data.

### 8.2.1 Ordered Logit

The Ordered Logit Model was initially selected as a conceptually appropriate choice for analyzing educational attainment, given that education levels follow a natural ranking. However, further diagnostic testing revealed that the proportional odds assumption was violated, as confirmed by the Likelihood Ratio Test for Proportional Odds. This assumption is fundamental to the Ordered Logit Model, as it requires that the effect of each predictor variable remains constant across all levels of the outcome variable. When this assumption is violated, the model imposes an incorrect structure on the data, leading to poor predictive performance and biased coefficient estimates. In fact, the model's poor classification performance persisted, with low test accuracy and a high Mean Absolute Error (MAE) of 1.7, indicating that on average, predictions were off by more than two education levels. The confusion matrix further confirmed that the model struggled to correctly classify middle education levels, likely because the effect of key predictors varies at different stages of educational progression, violating the core assumption of proportional odds. Given these limitations, we also tested the LogisticIT Model, an extension of logistic regression that allows for more flexibility in ordinal or multinomial classification settings. However, this approach did not yield meaningful improvements, with test accuracy remaining low and MAE still high (1.61). The violation of the proportional odds assumption suggests that educational attainment is not a simple ordinal process, but rather

one influenced by nonlinear and threshold-dependent factors. This means that the effect of predictors may differ depending on whether a student is transitioning from primary to secondary education or from secondary to higher education. Such variability cannot be adequately captured by standard Ordered Logit Models, making alternative approaches necessary. Given these findings, the next step involved shifting towards Multinomial Logistic Regression, which does not assume proportional odds and allows for more flexibility in modeling category-specific effects. Additionally, recognizing the limitations of parametric models in handling complex, nonlinear interactions, the focus shifted towards more robust non-parametric methods.

### 8.2.2 Multivariate Logistic Regression

**Parents' dataset**

The multinomial logistic regression model highlights the complex interplay of personal, socio-economic, and familial factors in shaping educational attainment among adoptees. The MNLogit was estimated using Maximum Likelihood Estimation (MLE) and successfully converged after 280 iterations, demonstrating numerical stability. The model significantly improves upon the null model (LL-Null: -1453.3 vs. Log-Likelihood: -867.47). The LLR p-value confirms strong statistical significance, while the HC3 robust covariance correction enhances reliability by accounting for heteroscedasticity. The mean absolute error (MAE) of 0.4542 on the training set and 0.3733 on the test set suggests that while the model makes relatively few large classification errors, some misclassification persists, particularly in underrepresented education categories. Although further experimentation with feature interaction terms and the removal of less significant predictors did not yield substantial improvements in model accuracy and ROC AUC metrics, suggesting that the relationships captured by the model are relatively stable, they helped to test theoretical assumptions further and derive more intricate connections between variables.

According to the model results for the parents' dataset, adoption age plays a key role, with later adoption increasing the likelihood of lower educational outcomes, while early adoption correlates with higher academic achievement. Schooling in the country of origin supports secondary education completion but does not necessarily lead to higher education. Health-related challenges significantly impact educational trajectories, with certified learning disabilities and disabilities under L.104/1992 strongly associated with lower attainment levels. Economic independence and employment negatively affect further education, suggesting financial constraints limit academic progression. Parental education has minimal influence, while having an adopted sibling positively correlates with higher academic success, possibly indicating

shared support mechanisms. Gender differences are present but weak, with females slightly more likely to complete secondary education. Unsurprisingly, academic performance, particularly school grades, remains the strongest predictor of university access.

For Secondary School Diploma (Class 1) and Professional Qualification Diploma (Class 2), older age at adoption significantly decreases the likelihood of attaining this qualification. Employment also negatively impacts this level, suggesting that individuals who enter the workforce early are less likely to complete secondary education. Conversely, entering the workforce could be merely a consequence of finishing the educational path early. The presence of serious disability certifications has by far the most significant effect at this level, reinforcing the idea that medical challenges present substantial barriers to education.

Regarding Upper Secondary School Diploma (Class 3), which permits university enrollment, adoption age has an even stronger negative effect together with employment. While international adoption emerges as a significant positive factor, suggesting that students with international backgrounds may be more likely to pursue 5-year high school which allows access to university. Higher maternal education also slightly increases the probability of achieving this diploma, highlighting the potential role of parental influence in shaping academic success.

For Post-Secondary Non-Tertiary Qualification (Class 4), health-related conditions have the most pronounced negative impact, severely reducing the likelihood of reaching this level. Notably, gender becomes a significant predictor, with females being more likely to achieve this level, possibly reflecting differences in career pathways or educational choices. Maternal education is a positive predictor, reinforcing the importance of parental academic background, together with international study experiences, suggesting that exposure to diverse learning environments supports specialized training. Lastly, having an adopted sibling shows borderline significance. Lastly, for University Degree (Class 5), adoption age has the strongest negative effect, highlighting that early adoption is crucial for long-term academic success. Now, international adoption, which was a positive predictor at the upper secondary level, now becomes a negative factor, reinforcing the idea that cultural and academic transitions may create barriers to higher education. Learning disability certifications also have a strong negative effect, indicating that students with additional educational needs struggle to reach university completion. However, having an adopted sibling significantly increases the probability of obtaining a university degree, suggesting that shared experiences and familial support may provide resilience in higher education. Namely, interaction term international (adoption * presence of an adopted sibling)

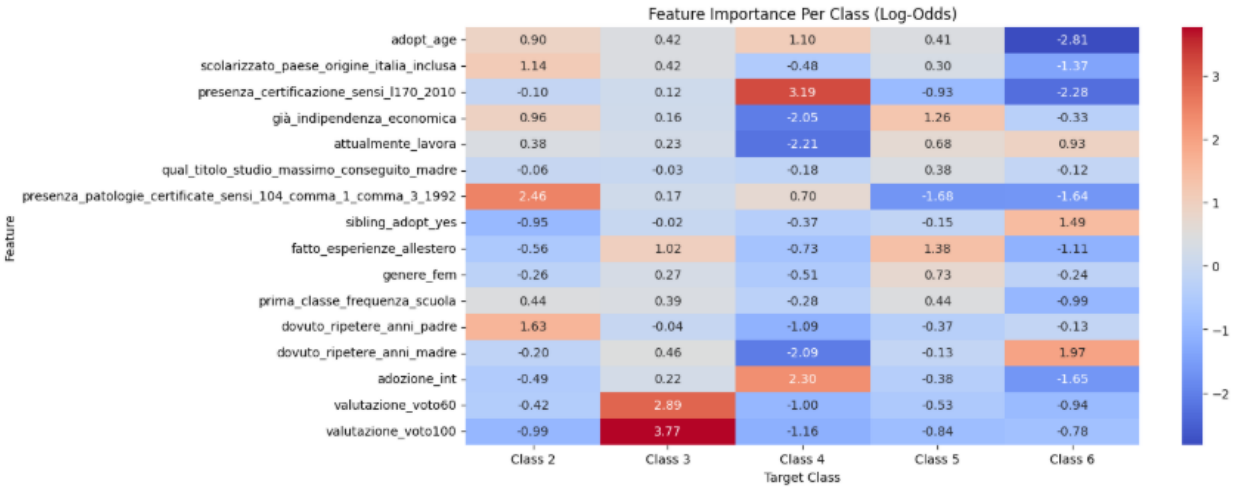



Figure 8.1: Log-Odds.

has had a strongly positive impact, highlighting that having a sibling with a similar experience might help non-Italian adoptees with achieving higher education levels.

Overall, the results indicate that older adoption age, employment, and health-related conditions pose significant barriers to higher educational attainment, while maternal education and sibling adoption can provide positive influences depending on the level of schooling. The findings highlight how structural and personal factors shape educational trajectories, with some predictors having differing effects depending on the stage of education.

**Children's dataset**

Similarly, the results from the multivariate logistic regression model on the children's responses provides a statistically significant, yet moderate model performance with the training accuracy of 75%, MAE of 0.6. Precision, recall, and F1-score vary across classes, with Class 3 achieving the highest recall (88%) and F1-score (74%), while Class 2 and Class 5 have lower recall values, indicating that the model struggles to correctly classify instances within these categories. Feature importance analysis sheds light on the key predictors driving the model's classification. The most influential variable is high school grade, which aligns with expectations, as academic performance is a strong determinant of educational success. Other significant features include teacher influence on school choice, first grade of school attended (which may be much higher for international adoptees), and maternal educational attainment.

For Professional Qualification Diploma (Class 2) and High School Diploma (Class 3), just as previously established, later adoption age and presence of health issues increases the likelihood of being attributed to this class. However, easier self-declared primary school experience decreases the likelihood, possibly indicating that those who

have felt like they excelled in the primary school environment may opt for different educational paths. International adoption also has a significant positive impact, suggesting that foreign-born adoptees may be more inclined toward vocational education, though this effect is moderated by gender. For Post-Secondary Non-Tertiary Qualification (Class 4), the findings are less conclusive. Unlike in Class 3, neither learning disability certifications nor health-related conditions significantly affect the likelihood of achieving this qualification. However, paternal education remains a significant positive predictor, reinforcing the idea that parental and maternal academic background continues to influence educational choices. The influence of early schooling experiences is also notable, with those who received strong foundational education in primary school being more likely to achieve this level.

Lastly, for University Degree (Class 5), older adoption age again negatively affects educational attainment, although the effect is weaker than in lower educational levels. International study experiences, however, have a strong positive effect, indicating that exposure to different educational systems and cultural environments enhances the probability of university completion or that it is generally available to a higher extent in university. Paternal education remains a key determinant, reinforcing the pattern seen in previous levels. Early primary school experiences also have a strong positive effect, suggesting that a strong educational foundation at an early age supports long-term academic success. Overall, the results indicate that older adoption

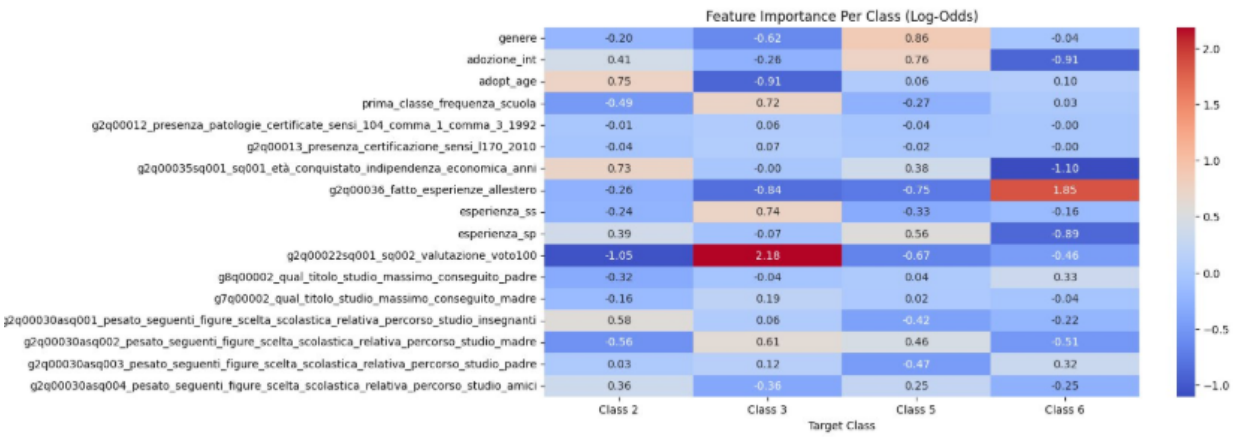



Figure 8.2: Log-Odds.

age generally reduces educational attainment, though the strength of this effect varies by level. Parental education, particularly paternal education, is consistently a strong positive predictor, reinforcing the role of family background in shaping academic trajectories. Practical work experience has mixed effects, reducing university

attainment but sometimes playing a role in vocational education. International adoption and international study experiences influence different levels differently, with vocational education benefiting from international adoption while university completion is positively influenced by international academic exposure. The role of teachers in educational decision-making appears to steer students away from post-secondary and university education, while maternal influence is a more consistent positive predictor. These findings suggest that early educational interventions, strong parental support, and international exposure can facilitate higher educational attainment, while delayed adoption and work experience often redirect students toward vocational paths.

Therefore, the logistic regression model provides a moderate level of accuracy in predicting educational attainment, with strong performance in certain classes but weaker classification in others. The feature importance analysis confirms that academic performance, parental education, and teacher influence are the strongest predictors, while factors such as health conditions and adoption status contribute minimally. Future model refinements could involve feature selection to remove non-informative variables, addressing class imbalance, and testing alternative classification approaches such as Random Forest or Gradient Boosting for improved predictive accuracy and generalization. In addition to these refinements, Structural Equation Modeling (SEM) can provide deeper insights by capturing latent relationships and indirect effects that logistic regression may overlook and provide a better outlook on predictors' significance.

## 8.3 Structural Equation Modeling (SEM)

SEM was employed for both datasets with the aim to explore whether variables which had rather negligible contribution to multivariate logistic regression results might turn out to have higher impact if paired with other aspects, as unlike traditional regression models, SEM accounts for indirect effects, measurement errors, and multiple interdependent relationships, providing a more holistic view of the factors influencing educational trajectories. This makes SEM especially valuable in analyzing adoption-related variables, socioeconomic conditions, and academic performance, where interactions and latent constructs might play a crucial role. The SEM analysis was carried out with the help of *semopy* package in Python, as outlined in the paper of Meshcheryakov G., Igolkina A. (2020). Notably, *semopy* offers several advantages over other SEM software, as it is optimized for Python, integrating seamlessly with machine learning workflows and data pipelines. One key advantage is its use of

automatic differentiation and modern optimization algorithms, allowing for faster and more accurate parameter estimation, particularly for large or complex models. Unlike *lavaan*, which relies on traditional maximum likelihood estimation, *semopy* supports flexible estimation methods and can handle categorical and continuous variables more efficiently. This is particularly important for the adoption survey datasets due to the presence of categorical and non-normally distributed variables in the adoption dataset, as the use of Diagonally Weighted Least Squares (DWLS) over Maximum Likelihood (ML) resulted in better convergence and overall model performance. This aligns with findings from Mîndrilă (2010), which suggest that DWLS is better suited for ordinal and non-normally distributed data. Additionally, semopy is built for scalability and computational efficiency, making it a better choice for integrating SEM within broader machine learning frameworks.

**Parents' Dataset**

The SEM results for the adults' dataset provide valuable insights into the relationships between personal, school, and family factors in shaping educational attainment. The inclusion of latent variables such as personal (comprising gender, international experiences, and adoption status), school (academic experiences), and family (parental education and household dynamics) ensures that complex relationships between observed and unobserved factors are considered. Additionally, the model includes direct effects for health-related challenges, acknowledging the significant role that disabilities play in educational outcomes. By allowing school and family to correlate, the model recognizes that these domains are not independent but influence each other in shaping an individual's academic trajectory. However, it is important to note that parental variables, such as education level and repeating school years, exhibited high multicollinearity during VIF testing, which distorted the SEM results and made their simultaneous inclusion unfeasible. This multicollinearity arose primarily because (1) mothers and fathers tended to have similar education levels, a common real-life pattern, and (2) school repetition was overwhelmingly rare, with the majority of responses being zeros, leading to an imbalanced distribution. To maintain model stability and interpretability, like in previous regression models, these variables were included separately for either the mother or the father rather than together, even though this choice was not theoretically ideal.

The following SEM model represents the relationships between family structure, personal factors, cognitive ability, and educational attainment:

$$\text{family\_structure} = \lambda_1 \cdot \text{sibling} + \lambda_2 \cdot \text{edu\_mother} + \lambda_3 \cdot \text{repeat\_father} + \epsilon_1,$$
$$\text{personal} = \lambda_4 \cdot \text{gender} + \lambda_5 \cdot \text{adopt\_age} + \lambda_6 \cdot \text{abroad} + \lambda_7 \cdot \text{adoption} + \epsilon_2,$$
$$\text{cognitive} = \lambda_8 \cdot \text{health\_issues} + \lambda_9 \cdot \text{certification} + \epsilon_3.$$

The latent variable correlations:

$$\text{personal} \sim \text{family\_structure}, \quad \text{cognitive} \sim \text{personal}.$$

Finally, the outcome equation:

$$\text{edu\_attainment} = \beta_1 \cdot \text{family\_structure} + \beta_2 \cdot \text{cognitive} + \beta_3 \cdot \text{personal} + \epsilon_4.$$

Fit indices suggest that the model is reasonably well-specified, with the chi-square statistic ($\chi^2 = 62$, p = 0.08) indicating that the model does not significantly deviate from the observed data, meaning it provides a good approximation of reality. The Comparative Fit Index (CFI) of 0.89 suggests a good fit, and Root Mean Square Error of Approximation of 0.03 is well within the acceptable range, indicating minimal unexplained variance and strong model stability.

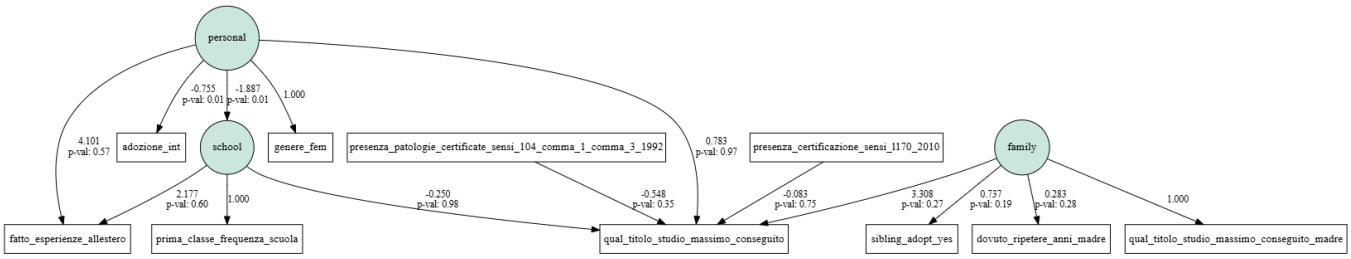


Figure 8.3: Best-performing SEM model - Parents' Dataset.

However, many of the estimated direct effects did not reach statistical significance, indicating that these factors may not strongly predict educational outcomes in adulthood when considered independently. Personal factors, including gender, international experiences, and adoption status, showed weak and non-significant effects on school outcomes and educational attainment. Similarly, school-related variables, such as first grade of insertion into the school system post-adoption, did not meaningfully impact educational attainment (p = 0.2). While family background appeared to have a relatively larger effect (Estimate = 3), it also did not reach statistical significance (p = 0.34), suggesting that while familial influences

are important, their direct impact on final educational outcomes may be moderated by other, unaccounted-for variables. A key finding, however, is the significant and negative effect of certified disabilities on educational attainment (Estimate = -0.6, $p < 0.001$), as previously seen in other tests. Furthermore, the correlation between school and disabilities was marginally significant ($p = 0.051$), reinforcing the idea that health-related challenges may indirectly affect school grades by influencing school experiences.

**Childrens' Dataset**

Similarly, the best model approach was to mix factors with individual variables that had the most success in terms of fit and interpretability. Baseline model that included all the variables as parts of factors following a theoretical approach resulted in an overly complex structure, which led to a convergence failure. Also, unlike in the parents' dataset where we had over 300 observations, childrens' data is more limited, with a sample size of a little over 100, which is a bit too small for SEM guidelines.

Overall, the SEM model which embedded both theory and preliminary assessment outcomes was structured as:

$$\begin{aligned}
\text{personal_factors} &= \lambda_1 \cdot \text{gender} + \lambda_2 \cdot \text{adoption} + \lambda_3 \cdot \text{adopt_age} + \epsilon_1, \\
\text{family_support} &= \lambda_4 \cdot \text{teachers} + \lambda_5 \cdot \text{mother} + \lambda_6 \cdot \text{friends} + \epsilon_2, \\
\text{education_exp} &= \lambda_7 \cdot \text{abroad_exp} + \lambda_8 \cdot \text{grades} + \epsilon_3.
\end{aligned}$$

The dependent variable equation:

$$\begin{aligned}
\text{edu_attainment} = \beta_1 \cdot \text{personal_factors} + \beta_2 \cdot \text{family_support} + \beta_3 \cdot \text{education_exp} \\
+ \beta_4 \cdot \text{econ_independence} + \beta_5 \cdot \text{health_issues} + \epsilon_4.
\end{aligned}$$

The latent variable correlations:

$$\begin{aligned}
\text{personal_factors} &\sim \text{econ_independence}, \\
\text{education_exp} &\sim \text{econ_independence} + \text{health_issues}.
\end{aligned}$$

The best strategy was to remove low-impact variables, even if it meant getting rid of a factor and using the most influential variable and an individual predictor. For instance, that was the case for disabilities, as learning disorders showed lower correlation, so keeping it along with another disability variable with a broader definition only distorted the model results. At the end, the following model structure

was the most successful, with factors like family support, educational experience, and personal information like sex, adoption type, and adoption age. The chi-square test of model fit ($\chi^2 = 28.47$, p = 0.493) is non-significant, indicating that there is no substantial discrepancy between the model-implied covariance matrix and the observed covariance matrix. This suggests that the hypothesized relationships among the latent constructs and observed variables sufficiently capture the underlying data structure. The Goodness-of-Fit Index of 0.81 and Adjusted Goodness-of-Fit Index of 0.7 suggest an overall good representation of the data, although the AGFI remains slightly below the ideal threshold of 0.8, indicating some room for refinement. The RMSEA below 0.05 indicates a close approximation of the population covariance structure.

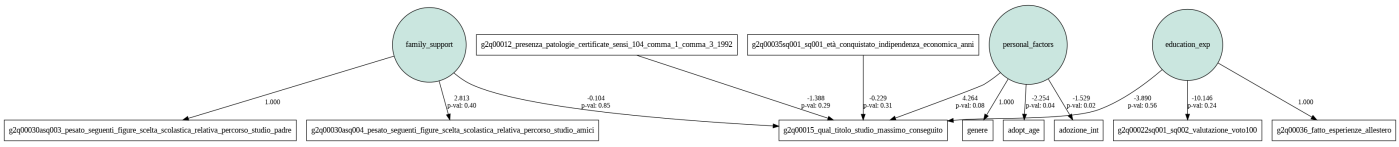



Figure 8.4: Best-performing SEM model - Children's Dataset.

The results from the structural equation model indicate several significant relationships among the latent constructs and observed variables. International adoption status ( p = 0.024) and age at adoption (p = 0.042) show significant negative associations with personal factors, which overall is a borderline significant factor for the target variable (p = 0.08). Additionally, a significant correlation exists between personal factors and economic independence (p = 0.04), indicating that personal characteristics may influence financial self-sufficiency. However, family support and education experience do not show significant direct effects on educational attainment, suggesting that other unobserved factors might be mediating these relationships. Overall, while the model confirms some expected relationships, such as the strong negative impact of disabilities on educational attainment, it also highlights the complex interplay of school, family, and personal factors. The lack of strong direct effects suggests that educational success in adulthood may be influenced by more nuanced, long-term interactions rather than isolated predictors, emphasizing the need for a more holistic approach in studying adoptees' academic trajectories.

## 8.4 Advanced Non-Parametric Methods

Eventually, non-parametric tests were employed using the upsampled training sets to uncover different patterns that may not be captured by Multinomial Logistic Regression and Structural Equation Modeling. Given their flexibility, non-parametric

approaches allow for detecting relationships without assuming linearity or distributional constraints, making them particularly useful in identifying hidden trends, interactions, and group differences that might be overlooked in traditional models. Additionally, these tests serve as a robust validation tool, helping to confirm or challenge the findings from multinomial logit and SEM by assessing whether the same predictors remain significant when analyzed without strict assumptions. This dual approach ensures a more comprehensive and unbiased understanding of the factors influencing educational attainment, increasing the reliability of the conclusions drawn. However, their interpretability is not as straightforward, as non-parametric models do not provide direct coefficient estimates or p-values, making it harder to quantify the exact influence of each variable (Derrick et al., 2020).

### 8.4.1 CatBoost Algorithm

CatBoost's main advantages are that it natively handles categorical data as well as its regularization techniques and bagging strategies which help reduce overfitting, making it well-suited for a small dataset with complex relationships. Hyperparameter tuning focused on maximizing classification performance while ensuring minority classes were detected. The final model used 2200 iterations, a tree depth of 3, and a learning rate of 0.005. The high number of iterations allowed for thorough learning, while the moderate tree depth controlled complexity and prevented overfitting. Lastly, the model was evaluated using the Total F1-score along with accuracy metrics to prioritize balanced classification and balance out the model fit.

Table 8.5: CatBoost Feature Importance Ranking - Parents' Dataset

| **Feature** | **Importance** |
|---|---|
| valutazione_voto100 | 30.033879 |
| adopt_age | 20.932221 |
| attualmente_lavora | 8.402160 |
| qual_titolo_studio_massimo_conseguito_madre | 7.883730 |
| già_indipendenza_economica | 7.420236 |
| prima_classe_frequenza_scuola | 6.471209 |
| presenza_certificazione_sensi_l170_2010 | 4.602000 |
| dovuto_ripetere_anni_madre | 3.566802 |
| adozione_int | 2.723282 |
| presenza_patologie_certificate_sensi_104_comma | 2.451499 |
| scolarizzato_paese_origine_italia_inclusa | 1.733067 |
| genere_fem | 1.656281 |
| sibling_adopt_yes | 0.960956 |
| fatto_esperienze_allestero | 0.734461 |
| dovuto_ripetere_anni_padre | 0.428217 |

In the parents dataset, the tuned CatBoost model achieved slightly better results thanks to the larger sample size for training, further enhanced by the upsampling techniques. Overall, it had a cross-validated accuracy of 76%, improving upon previous models. The difference between training and test MAE (0.45 vs. 0.6) remained within an acceptable range without severe overfitting, reinforcing the model's generalization capability. The feature importance analysis highlights academic performance and employment as the strongest predictors of educational outcomes. Economic independence, learning disability certifications, early schooling, and maternal education contribute moderately, reflecting socioeconomic influence. The results confirm that academic history, adoption-related factors, and socioeconomic conditions drive classification, suggesting future refinements should focus on feature interactions to improve predictive accuracy.

Applying a similarly structured model to the children's dataset yielded better results, as there were no signs of overfitting behavior, with the overall model cross-validated accuracy of 75%. The results highlight age of economic independence as the most influential predictor, followed by academic performance and international experiences, suggesting that self-sufficiency, grades, and exposure to foreign environments significantly impact educational outcomes. Parental influence on decision-making and adoption age also play a role, though slightly less than in the parents' dataset. Parental education and specific social experiences contribute, while factors like gender and international adoption have minimal impact. Unlike the parents' dataset, certifications for disabilities hold little significance, suggesting that children perceive independence and personal experiences as more critical to their educational paths. However, this can merely be a reflection of the fact that despite upsampling some groups are underrepresented in a given class, which prompts the model to choose a more arbitrary criteria. So, although CatBoost results confirm some of the values' importance, it is important to underline the limitations that come along with it.

### 8.4.2 Support Vector Machine

The Support Vector Machine model with RBF kernel, optimized through grid search, exhibited strong predictive performance on the parents' dataset, achieving a train ROC AUC of 0.92 and a test ROC AUC of 0.86, with a smaller overfitting gap of 0.06. However, the overfitting gap of 0.08 was notably higher than that observed in the children's dataset, where SVM achieved a train ROC AUC of 0.82 and a test ROC AUC of 0.85. This suggests that while the SVM model trained on the parents' data was highly effective in capturing patterns within the training set, it struggled to generalize as effectively as the children's model due to the lack of class 4 samples

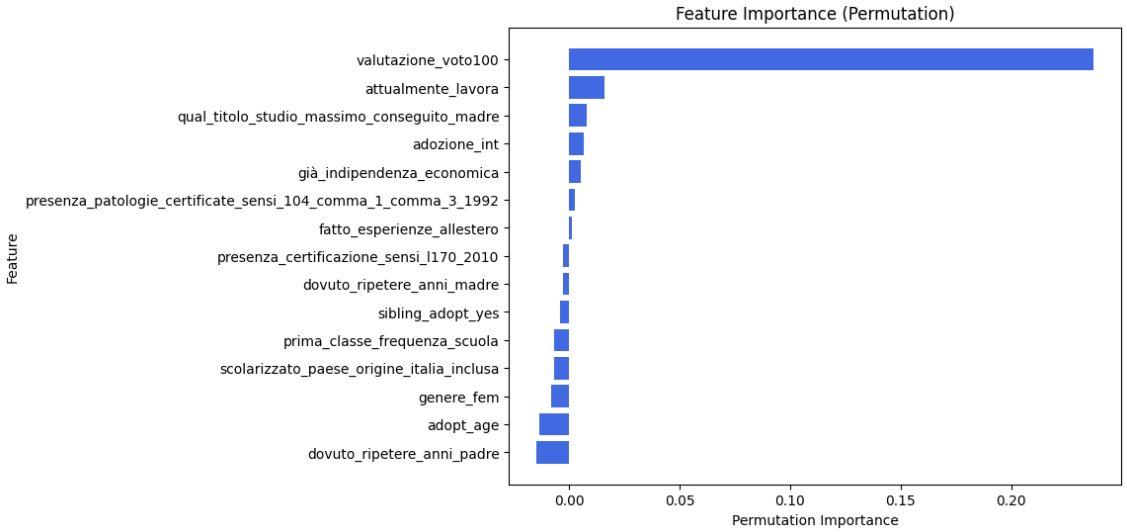
Table 8.6: CatBoost Feature Importance Ranking - Children's Dataset

| **Feature** | **Importance** |
|---|---|
| g2q00035sq001_sq001_età_conquistato_indipendenza | 21.034189 |
| g2q00022sq001_sq002_valutazione_voto100 | 18.048485 |
| g2q00036_fatto_esperienze_allestero | 12.744854 |
| g2q00030asq002_pesato_seguenti_figure_scelta_scolastica_madre | 6.541691 |
| g2q00030asq003_pesato_seguenti_figure_scelta_scolastica_altro | 6.388369 |
| adopt_age | 6.099850 |
| esperienza_sp | 4.580856 |
| g7q00002_qual_titolo_studio_massimo_conseguito_madre | 4.570136 |
| g8q00002_qual_titolo_studio_massimo_conseguito_padre | 3.701021 |
| prima_classe_frequenza_scuola | 2.986456 |
| esperienza_ss | 2.984225 |
| g2q00030asq001_pesato_seguenti_figure_scelta_scolastica_insegnanti | 2.797198 |
| g2q00030asq004_pesato_seguenti_figure_scelta_scolastica_amici | 2.681462 |
| genere | 2.501719 |
| adozione_int | 2.318539 |
| g2q00012_presenza_patologie_certificate_sensi_104_comma | 0.020953 |
| g2q00013_presenza_certificazione_sensi_l170_2010 | 0.000000 |

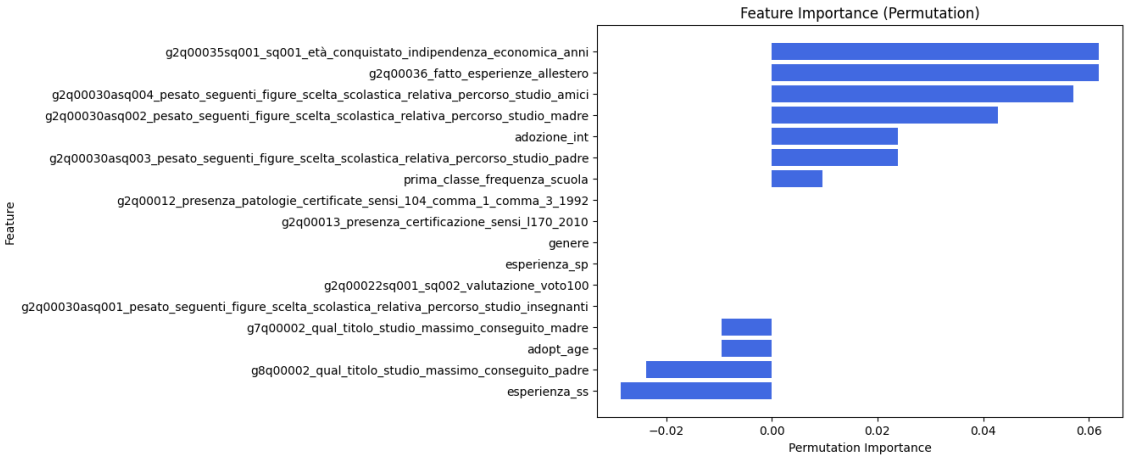
in the test dataset.

Both datasets highlighted key predictors of educational attainment, with age at adoption, international adoption status, and learning disability certifications emerging as the most influential variables. These findings reinforce the idea that structural and categorical attributes, particularly those related to early childhood experiences and educational challenges, play a defining role in academic trajectories. However, while both models identified learning disabilities as a crucial determinant, the parents' dataset placed greater emphasis on employment status, suggesting that financial independence and work commitments exert a stronger influence on long-term educational attainment. This contrasts with the children's dataset, where the impact of early-life experiences, including adoption-related factors, appeared more pronounced.

Despite its strengths in handling high-dimensional data, SVM's reliance on hyperplane-based classification presents limitations, particularly in dealing with class imbalances, which likely contributed to the misclassification of less frequent educational categories. This was evident in the lower recall for underrepresented groups. Compared to CatBoost, which better captures intricate feature interactions through gradient boosting, SVM operates with a more rigid, rule-based approach, excelling at making structural distinctions but less effective at uncovering deeper, nonlinear relationships. This contrast underscores the trade-offs between different modeling




(a) SVM Feature Importance - Parents' Dataset.




(b) SVM Feature Importance - Children's Dataset.

Figure 8.5: Comparison of CatBoost Feature Importance for Parents' and Children's Datasets.

approaches in understanding educational outcomes, where balancing interpretability, generalizability, and predictive power remains a central challenge.

# 9 Conclusions

This study employed a combination of parametric and non-parametric methods to explore the educational trajectories of adopted children, aiming to balance predictive accuracy with interpretability, and achieving interesting results. By integrating multinomial logistic regression, structural equation modeling, and machine learning approaches such as CatBoost and SVM, the analysis provided a comprehensive perspective on how academic performance, adoption-related variables, socioeconomic factors, and learning disability certifications influence educational outcomes. While non-parametric ML models offered structured insights into variable interactions, they struggled with class imbalance and tended to favor continuous predictors, making them less effective in capturing categorical complexities and feature dependencies. This limitation was particularly evident in the adults' dataset, where the lack of Class 4 samples led to overfitting, whereas the children's dataset exhibited more stable model performance with hyperparameter tuning.

Firstly, non-parametric statistical tests such as the Chi-Square and Kruskal-Wallis tests were instrumental in the early stages of analysis, providing an initial outlook on variable relationships and helping to pre-select significant predictors. These tests confirmed that adoption-related factors, disabilities, and parental education played significant roles in shaping educational outcomes. However, their interpretability was inherently limited, as they did not quantify effect sizes or facilitate interaction analysis apart from inference potential available thanks to combining test results with contingency tables outlook. While they provided a foundation for subsequent modeling, their results needed to be supplemented by more advanced statistical and machine learning techniques.

Secondly, Multinomial logistic regression highlighted key relationships between predictors and educational attainment, confirming that older adoption age, employment, and health-related conditions significantly reduced the likelihood of achieving higher education levels. The presence of learning disability certifications posed substantial barriers at lower education levels, reinforcing the importance of specialized educational support. International adoption played a complex role, initially emerging as a positive factor in attaining an upper secondary school diploma but later becoming a negative predictor for university completion, suggesting that cultural and academic transitions may create challenges in higher education. The analysis further revealed that maternal education had a consistent, albeit modest, positive

influence, particularly in promoting pathways to higher education. Interestingly, the presence of an adopted sibling emerged as a resilience factor, particularly at the university level, where the interaction between international adoption and sibling adoption showed a strongly positive effect.

Structural equation modeling further supported these findings by identifying the broader structural relationships among adoption-related factors, disabilities, and socioeconomic influences. However, attempts to combine individual predictors into latent factors did not reveal any entirely new trends beyond reinforcing the critical role of personal characteristics such as adoption age, international adoption status, and the presence of disabilities. While SEM proved useful for mapping theoretical relationships, its reliance on strong distributional assumptions made it less practical for this dataset.

Regarding Machine learning approaches, CatBoost and SVM offered alternative perspectives on educational attainment, but their effectiveness varied in addressing class imbalance and interpretability challenges. CatBoost performed well in handling categorical variables and detecting non-linear interactions, but its ability to mitigate class imbalance was limited by the underlying data distribution, meaning that under-represented categories, such as Class 4 in the adults' dataset, remained difficult to classify accurately. SVM, on the other hand, struggled more with class imbalances, likely favoring majority classes in its predictions rather than prioritizing specific structural attributes like learning disabilities or adoption status. While both models demonstrated strong classification performance, their interpretability remained an issue, as they lacked direct coefficient estimation and statistical significance testing. However, CatBoost's SHAP values provided a degree of insight into feature importance, showing that academic performance, economic stability, and adoption-related factors played key roles in predicting outcomes. Despite differences in approach, some alignment with parametric models was observed, particularly in identifying adoption age, parental education, and learning disabilities as significant factors.

Each method presented distinct trade-offs, with parametric models offering interpretability and formal significance testing, while non-parametric models demonstrated superior predictive accuracy and flexibility in handling categorical data. CatBoost emerged as the most effective approach for capturing complex relationships, though its interpretability challenges made it less suitable for direct statistical inference. Logistic regression and SEM, despite their advantages in explaining structured relationships, struggled with class imbalances and rigid modeling assumptions. SVM, while capable of managing high-dimensional data, proved less suited for this context due to classification biases stemming from unbalanced sample distributions. These

limitations highlight the need for a more integrated methodological approach that combines statistical inference with machine learning to leverage the strengths of both paradigms.

The results ultimately suggest that older adoption age, employment, and health-related conditions act as substantial barriers to higher educational attainment, while maternal education, sibling adoption, and international experiences can provide positive influences depending on the stage of education. The patterns observed across different methods reinforce the notion that educational trajectories among adopted children are shaped by a combination of personal, structural, and socioeconomic factors that are, however, unique and require a tailored approach. Given the constraints of the dataset, particularly the imbalance across educational categories, reapplying parametric models with an expanded dataset would likely enhance the reliability of the findings.

## 9.1 Limitations and Future Work

Finally, it is important to acknowledge several limitations that impacted the robustness of the results. One of the most significant challenges was the small sample size, which restricted the statistical power of complex models and made it difficult to derive broad conclusions. The limited dataset particularly affected models that required large sample sizes for accurate estimation, such as Structural Equation Modeling (SEM) and machine learning approaches, leading to overfitting in some cases.

Another key limitation was the difficulty in classifying university graduates and other underrepresented categories due to class imbalances. Since the majority of respondents fell into the middle education categories, predictions for higher education levels were less reliable. This issue was especially prevalent in non-parametric models like CatBoost and SVM, where underrepresented groups were not adequately learned by the model.

Additionally, the dataset was predominantly categorical and non-normally distributed, making it less suitable for traditional parametric tests. Many standard statistical techniques, such as Ordinary Least Squares (OLS) regression, rely on normality assumptions that were not met in this study. As a result, non-parametric and robust statistical methods had to be employed to ensure meaningful insights while adjusting for heteroscedasticity and class imbalance.

To address these limitations and improve future research, several key areas for further investigation are proposed:

**1. Increase Sample Size**: Expanding the dataset, particularly for the children's survey, would enhance statistical power and improve model accuracy. A larger sample size would allow for more reliable subgroup analyses and mitigate class imbalance issues in higher education categories.

**2. Longitudinal Study**: Conducting a longitudinal study would provide deeper insights into the long-term educational trajectories of adopted children. Tracking participants over time would help identify causal relationships between early-life factors and later academic success, improving the validity of findings.

**3. Refined Data Collection**: Future studies should incorporate more nuanced questions to capture additional contextual factors influencing educational outcomes. Specifically, refining survey instruments to include indicators aligned with Self-Determination Theory could help explore the role of autonomy, competence, and relatedness in shaping academic achievement among adoptees.

By addressing these areas, future research can offer more comprehensive and generalizable insights into the educational experiences of adopted children, ultimately contributing to improved policy interventions and support mechanisms tailored to their needs.

# 10 Appendix

## 10.1 Contingency Tables - Parents' Dataset

Table 10.1: Highest Educational Qualification by Region of Origin

| Origin Region | No Title / Primary | Lower Secondary | Professional Diploma | Upper Secondary | Post-Secondary (IFTS) | University Degree | Postgraduate |
|---|---|---|---|---|---|---|---|
| Africa | 0 | 1.0 | 0 | 16.0 | 1.0 | 2.0 | 1.0 |
| Asia | 0 | 3.0 | 3.0 | 42.0 | 2.0 | 11.0 | 2.0 |
| East Europe | 0 | 8.0 | 27.0 | 73.0 | 4.0 | 6.0 | 0 |
| Italy | 0 | 10.0 | 7.0 | 47.0 | 1.0 | 10.0 | 3.0 |
| Latin America | 1.0 | 12.0 | 13.0 | 44.0 | 1.0 | 12.0 | 3.0 |
| Other | 0 | 0 | 0 | 4.0 | 1.0 | 2.0 | 0 |

Table 10.2: Contingency Table: Economic Independence

| Education Level | Not Independent (0) | Independent (1) |
|---|---|---|
| 2 | 51 | 28 |
| 3 | 176 | 37 |
| 4 | 8 | 1 |
| 5 | 17 | 23 |
| 6 | 4 | 5 |

Table 10.3: Contingency Table: Employment Status

| Education Level | Unemployed (0) | Employed (1) |
|---|---|---|
| 2 | 32 | 47 |
| 3 | 142 | 70 |
| 4 | 6 | 2 |
| 5 | 13 | 29 |
| 6 | 2 | 7 |

Table 10.4: Contingency Table: Certified Disabilities (Law 104)

| Education Level | No (0) | Yes (1) |
|---|---|---|
| 2 | 60 | 18 |
| 3 | 194 | 22 |
| 4 | 7 | 2 |
| 5 | 40 | 0 |
| 6 | 8 | 0 |

## 10.2 Contigency Tables - Childrens' Dataset

Table 10.5: Contingency Table: Sibling Adopted

| **Education Level** | **No (0)** | **Yes (1)** |
|---|---|---|
| 2 | 63 | 22 |
| 3 | 182 | 44 |
| 4 | 5 | 5 |
| 5 | 24 | 19 |
| 6 | 3 | 6 |

Table 10.6: Contingency Table: Experience Abroad

| **Education Level** | **No (0)** | **Yes (1)** |
|---|---|---|
| 2 | 79 | 6 |
| 3 | 170 | 56 |
| 4 | 7 | 3 |
| 5 | 28 | 15 |
| 6 | 7 | 2 |

Table 10.7: Contingency Table: Gender

| **Education Level** | **Male (0)** | **Female (1)** |
|---|---|---|
| 2 | 55 | 30 |
| 3 | 132 | 94 |
| 4 | 6 | 4 |
| 5 | 16 | 27 |
| 6 | 3 | 6 |

Table 10.8: Contingency Table for Educational Attainment and Mother's Influence on School Choice

| **Educational Attainment** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 2 | 0 | 4 | 5 | 0 |
| 2 | 0 | 3 | 7 | 0 | 4 | 0 | 0 |
| 3 | 1 | 6 | 14 | 0 | 6 | 2 | 0 |
| 4 | 1 | 4 | 20 | 2 | 11 | 8 | 1 |

Table 10.9: Contingency Table for Educational Attainment and Teachers' Influence on School Choice

| **Educational Attainment** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 11 | 1 | 7 | 3 | 0 |
| 2 | 1 | 5 | 6 | 0 | 10 | 4 | 1 |
| 3 | 3 | 7 | 20 | 1 | 5 | 4 | 0 |
| 4 | 0 | 2 | 6 | 0 | 3 | 4 | 0 |

Table 10.10: Multinomial Logistic Regression Results - Parents

| Variable | Category 3 | | | Category 4 | | | Category 5 | | | Category 6 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Coef | Std Err | p-val | Coef | Std Err | p-val | Coef | Std Err | p-val | Coef | Std Err | p-val |
| Intercept | -0.4910 | 0.558 | 0.379 | -8.4058 | 0.888 | 0.000 | -2.8745 | 0.675 | 0.000 | -5.0374 | 1.354 | 0.000 |
| adopt_age | -0.7481 | 0.205 | 0.000 | -1.9324 | 0.242 | 0.000 | -1.2751 | 0.242 | 0.000 | -4.9679 | 0.595 | 0.000 |
| certification | 0.9505 | 0.586 | 0.105 | 2.8604 | 0.633 | 0.000 | -0.8765 | 0.828 | 0.290 | 0.0000 | 0.846 | 1.000 |
| currently_working | -1.1966 | 0.321 | 0.000 | -3.5259 | 0.372 | 0.000 | 0.4189 | 0.281 | 0.136 | 0.5832 | 0.527 | 0.268 |
| mother_education | 0.1269 | 0.093 | 0.174 | 0.3729 | 0.104 | 0.000 | 0.4816 | 0.097 | 0.000 | 0.0875 | 0.169 | 0.604 |
| health_issues | -1.5497 | 0.566 | 0.006 | -0.3711 | 0.432 | 0.391 | -17.0618 | 2.473 | 0.000 | 0.0000 | 1.862 | 1.000 |
| first_school_year | -0.0514 | 0.165 | 0.755 | 0.1630 | 0.210 | 0.438 | 0.2653 | 0.203 | 0.191 | -0.7682 | 0.572 | 0.179 |
| repeated_years_mother | 0.1095 | 0.405 | 0.787 | -13.4003 | 2.883 | 0.000 | -0.4471 | 0.431 | 0.299 | 1.5296 | 0.696 | 0.028 |
| gender | 0.6474 | 0.655 | 0.323 | 0.0000 | 0.974 | 1.000 | 1.1832 | 0.638 | 0.064 | 0.6100 | 0.737 | 0.408 |
| adoption | 0.4676 | 0.448 | 0.297 | 7.5359 | 0.798 | 0.000 | 0.6889 | 0.526 | 0.190 | -5.6075 | 0.554 | 0.000 |
| gender:adoption | -0.0920 | 0.730 | 0.900 | 0.1354 | 1.029 | 0.895 | -0.1997 | 0.708 | 0.778 | -0.2010 | 0.922 | 0.827 |
| sibling_adopted | -0.2897 | 0.922 | 0.753 | 0.0000 | 1.497 | 1.000 | 0.8996 | 0.818 | 0.271 | 3.0399 | 1.069 | 0.000 |
| adoption:sibling_adopted | 1.0492 | 0.999 | 0.294 | 0.6985 | 1.550 | 0.652 | -0.1510 | 0.918 | 0.869 | 13.6905 | 1.260 | 0.000 |

Table 10.11: Multinomial Logit Model Results for Educational Attainment - Children

| Variable | Category 3 | | | Category 5 | | | Category 6 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Coef | Std Err | p-val | Coef | Std Err | p-val | Coef | Std Err | p-val |
| Intercept | -5.582 | 1.602 | 0.000 | -1.674 | 1.107 | 0.130 | -0.137 | 1.423 | 0.924 |
| adopt_age | -1.570 | 0.619 | 0.011 | -0.591 | 0.296 | 0.046 | -0.673 | 0.468 | 0.150 |
| health_issues | 5.745 | 1.489 | 0.000 | 0.000 | 0.494 | 1.000 | 0.000 | 1.338 | 1.000 |
| certification | 4.936 | 1.199 | 0.000 | 0.000 | 0.499 | 1.000 | 0.000 | 0.792 | 1.000 |
| abroad_exp | 0.171 | 0.867 | 0.844 | -0.014 | 0.742 | 0.984 | 3.012 | 0.863 | 0.000 |
| adoption | 7.029 | 2.594 | 0.007 | 1.632 | 1.588 | 0.304 | 0.714 | 2.626 | 0.786 |
| gender | -0.207 | 0.840 | 0.805 | 0.886 | 0.765 | 0.247 | 0.583 | 0.732 | 0.426 |
| gender:adoption | -2.253 | 1.399 | 0.107 | 0.948 | 1.380 | 0.492 | -0.373 | 1.956 | 0.849 |
| first_school_year | 1.544 | 0.541 | 0.004 | 0.435 | 0.373 | 0.244 | 0.626 | 0.446 | 0.161 |
| adoption:first_school | -1.816 | 0.714 | 0.011 | -0.707 | 0.476 | 0.138 | -0.926 | 0.534 | 0.083 |
| teachers_influence | -0.704 | 0.358 | 0.049 | -0.852 | 0.331 | 0.010 | -0.928 | 0.383 | 0.015 |
| mother_influence | 1.091 | 0.293 | 0.000 | 0.773 | 0.275 | 0.005 | 0.000 | 0.428 | 1.000 |

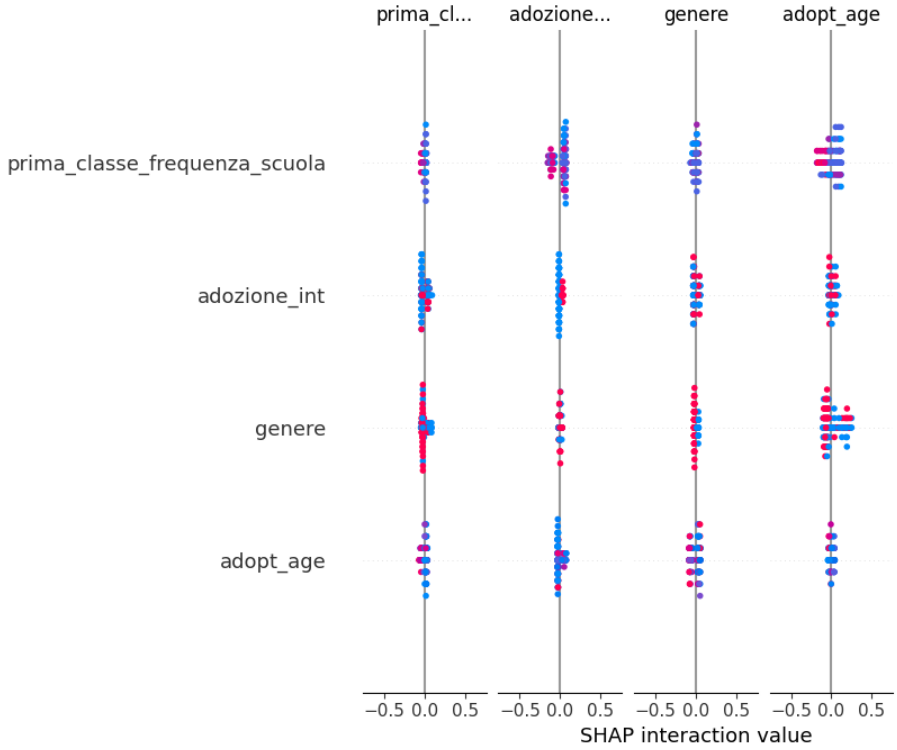



Figure 10.1: CatBoost SHAP - Children's Dataset.The SHAP interaction plot visualizes how pairs of features interact to influence the model's predictions, with the x-axis representing SHAP interaction values and the y-axis listing the interacting features. The color gradient (blue to red, low to high) represents the values of one of the interacting features, highlighting variation in influence across different observations. Features like 'adopt age' and 'adozione int' exhibit stronger interactions, indicated by the spread of points, while others, such as 'prima classe frequenza scuola', show a more centralized distribution, suggesting weaker or more stable effects in relation to other variables.

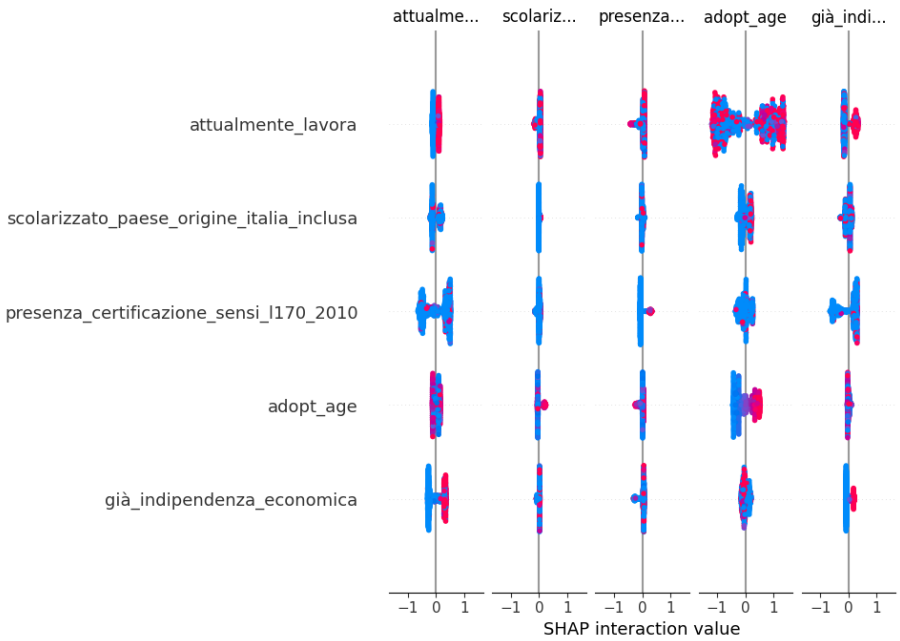



Figure 10.2: CatBoost SHAP - Parents' Dataset.Similarly to the previous graph, features like 'adopt age' and 'indipendenza economica' show strong interactions, as the points are wide spread, whereas others, such as 'presenza certificazione sensi l170 2010', have more compact distributions, indicating weaker effects.

# 11 Bibliography


Barcons, N., Abrines, N., Brun, C., Sartini, C., Fumadó, V., Marre, D., & Fornieles, A. (2014). Social relationships in children from intercountry adoption. *Children and Youth Services Review, 42*, 40-46. https://doi.org/10.1016/j.childyouth.2014.03.013

Balenzano, C., & Coppola, R. (n.d.). A retrospective study on adoptive parenthood in the Italian context.

Cai, Commissione Adozioni Internazionali. (2023). *Summary Report CAI 2023*. https://www.commissioneadozioni.it/media/iy3py3j0/summary-report-cai-2023_eng.pdf

Caceres, Child-2024. (n.d.). School victimization and psychosocial adjustment among Eastern European adopted adolescents across Europe.

CatBoost Team. (2024). *CatBoostClassifier Python Reference*. CatBoost AI Documentation. https://catboost.ai/docs/en/concepts/python-reference_catboostclassifier

Curran, P. (n.d.). Exploring the fit of structural equation models: Tests of significance and goodness-of-fit measures.

Derrick, B., White, P., & Toher, D. (2020). Parametric and non-parametric tests for the comparison of two samples which both include paired and unpaired observations. *University of the West of England*.

Dhaene, S., & Rosseel, Y. (2022). Resampling-based bias correction for small sample SEM. *Structural Equation Modeling: A Multidisciplinary Journal*.

Ferritti, M., Guerrieri, A., Mattei, L. (2001). The educational choices of adopted students. INAPP, University of L'Aquila, ANPAL.

Hoshino, T., & Bentler, P. M. (2013). Bias in factor score regression and a simple solution. *National Center for Biotechnology Information (NCBI)*. https://pmc.ncbi.nlm.nih.gov/articles/PMC3773873

Howard, J. A., Smith, S. L., & Ryan, S. D. (2004). A comparative study of child welfare adoption disruption in the U.S. *Journal of Child Welfare, 7*(3), 3-30. https://doi.org/10.1300/J145v07n03_01

Juffer, F., & Van IJzendoorn, M. H. (2005). Behavior problems and mental health referrals of international adoptees: A meta-analysis. *JAMA Pediatrics, 159*(5), 527-534.

Kong, S., Ahn, D., Kim, B., Srinivasan, K., Ram, S., Kim, H., Hong, A., Kim,

J., Cho, N., & Shin, C. (2020). A novel fracture prediction model using machine learning in a community-based cohort. *JBMR Plus, 4*(3), e10337. https://doi.org/10.1002/jbm4.10337

Lorenzini, S. (2018). Adozione internazionale: multiculturalità nell'identità? Una lettura educativa e interculturale. In M. Di Mauro B. Gehrke (Eds.), Multicultural identities: Challenging the sense of belonging (pp. 163-184). Colle Val d'Elsa (SI): Fondazione Intercultura. https://hdl.handle.net/11585/634170.

Mîndrilă, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society (IJDS, 1*(1), 60–66.

Meshcheryakov, G., & Igolkina, A. A. (2019). Semopy: A Python package for structural equation modeling. *ArXiv*. https://arxiv.org/abs/1909.10758

Meshcheryakov, G., Igolkina, A. A., Samsonova, M. G. (2021). SEMOPY 2: A structural equation modeling package with random effects in Python. *ArXiv*. https://doi.org/10.48550/arXiv.2106.01140

Muntean, A. (2019). Late-adopted children grown up: A long-term longitudinal study on attachment patterns of adolescent adoptees and their adoptive mothers. *ResearchGate*. https://www.researchgate.net/publication/330856140_Late-adopted_children_grown_up_a_long-term_longitudinal_study_on_attachment_patterns_of_adolescent_adoptees_and_their_adoptive_mothers

Organisation for Economic Co-operation and Development (OECD). (n.d.). OECD family database. OECD. https://www.oecd.org/en/data/datasets/oecd-family-database.html

Ostroumova, L., Gusev, G., Vorobev, A., Dorogush, A., & Gulin, A. (2017). CatBoost: Unbiased boosting with categorical features.

Pace, C. S., et al. (2008). Behavioral and emotional problems among Italian international adoptees and non-adopted children: Fathers' and mothers' reports. *ResearchGate*. https://www.researchgate.net/publication/23196307_Behavioral_and_Emotional_Problems_Among_Italian_International_Adoptees_and_Non-Adopted_Children_Father's_and_Mother's_Reports

Palacios, J., Brodzinsky, D. (2010). Adoption research: Trends, topics, outcomes. International Journal of Behavioral Development, 34(3), 270-284. https://doi.org/10.1177/0165025410362837

Russo, N. (2015). Modelli a Equazioni Strutturali e Alberi Decisionali. Applicazione alla Previsione della Rinuncia agli Studi nella Laurea Triennale in Matematica. Unpublished manuscript

Schermelleh-Engel, K., & Moosbrugger, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Goethe University, Frankfurt.*

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66*(4), 563–575. https://doi.org/10.1007/BF02296196

Van Lissa, C. J., Garnier-Villarreal, M., & Anadria, D. (2023). Recommended practices in latent class analysis using the open-source R-package tidySEM. *Structural Equation Modeling: A Multidisciplinary Journal*. https://www.tandfonline.com/doi/full/10.1080/10705511.2023.2250920

Welsh, J. A., Viana, A. G., Petrill, S. A., & Mathias, M. D. (2019). Interventions for internationally adopted children and families: A review of the literature. *Child Abuse & Neglect, 88*, 1-12. https://doi.org/10.1016/j.chiabu.2018.10.021