Dean&Francis
ISSN 2959-6157

Comparative Analysis of Statistical Methods for Investigating Risk Factors of Adolescent Depression

Yixuan Wu

The University of British Columbia, Vancouver, V6T-1Z4, Canada
yixuanwu075@gmail.com

Abstract:

Depression often makes it difficult for individuals to form normal interpersonal relationships and may even lead to suicidal tendencies. Adolescents are in a critical period of developing psychological independence, with the prevalence of depression showing an increasing trend, making it a major global public health challenge. To achieve early identification and effective intervention, it is particularly important to employ appropriate statistical models to explore factors associated with adolescent depression. This article reviews the applications of multiple linear regression, logistic regression, linear mixed-effects models, and mediation models. The literature examines the specific problems addressed and the types of data suitable for each model in adolescent depression research. Based on the theoretical features of these models, comparisons are made to clarify how to choose the appropriate model under different research scenarios. Findings indicate that logistic regression is suitable for binary outcomes such as "whether an individual has depression"; multiple linear regression is often used for continuous variables such as depression scores; linear mixed-effects models are more appropriate for long-term follow-up studies of the same population or research involving individual variations; while mediation models are applicable when researchers aim to understand the mechanisms through which a certain factor influences the outcome.

Keywords: Statistical methods; Adolescent depression; Comparative analysis

1. Introduction

Depression is one of the most common mental health disorders in the world; the prevalence of it among adolescents has become a significant public health concern. In recent years, the prevalence of adolescent depression has been steadily increasing [1]. Adolescence represents a key transitional period between

ISSN 2959-6157

childhood and adulthood. The characteristic of it is significant physical, psychological, and emotional changes, enhancing adolescents' sensitivity and reactivity to stressors [2]. Boredom, difficulties of emotional regulation, and the decline of coping abilities increased the risk of having depression [3]. The severe negative consequences of depression include social isolation, suicide, and impaired functioning [4]. Early detection of depressive symptoms in adolescents is particularly important because delayed treatment can lead to worsening symptoms [5]. Accurate identification of the causes of adolescent depression is the foundation for developing treatment strategies.

Researchers have conducted extensive studies on factors influencing adolescent depression, examining multiple domains including family factors, school environments, social support, gender, and others. These studies employed a wide range of statistical methods, such as using mean and standard error to assess overall data patterns, and employing linear regression and other techniques to examine relationships between predictors and outcomes [6-9]. However, what truly determines the validity and generalizability of research is often the fit between the model and the data structure, rather than merely focusing on the importance of individual factors. Adolescent depression research typically involves high-dimensional, multimodal information (e.g., scale scores, diagnostic categories, event timing measures, repeated observations, and nested group structures) and is often accompanied by statistical challenges such as multicollinearity, missing data, skewed distributions, and category imbalance. In this context, how to select appropriate models based on research questions and data characteristics, and how to clearly report assumptions, estimates, and uncertainties, become critical factors influencing the credibility, interpretability, and reproducibility of results. The lack of systematic comparisons of commonly used models under different conditions also limits the availability of practical methodological guidance [10].

A review of numerous statistical studies on depression risk factors reveals that multiple linear regression, logistic regression, mediation models, and linear mixed-effects models are widely used in adolescent depression research. Linear regression is commonly used to analyze linear relationships between continuous outcome variables (e.g., depression scale scores) and multiple risk factors [11]. For binary outcomes (e.g., occurrence of depressive episodes), logistic regression is more appropriate. It calculates odds ratios to assess the relative impact of different factors [12]. Mediation models introduce a mediating variable to analyze how independent variables influence dependent variables through intermediate mechanisms [13]. Linear mixed-effects models demonstrate particular adaptability when handling repeated measures or hierarchical data (e.g., individuals nested within families or schools) [14].

In this context, this review focuses on four statistical methods widely used in epidemiological and mental health research: logistic regression, multiple linear regression, linear mixed-effects models, and mediation models. The primary objective of the study is to conduct a systematic review of peer-reviewed literature published between 2020 and 2025 on adolescent depression, examining the specific applications of these models in different research contexts. This paper will compare the applicability of the four models across various aspects, including the type of dependent variables (e.g., continuous scale scores, binary outcomes), data structure (cross-sectional, longitudinal, or hierarchical/repeated measures), and summarize differences in categorical variable coding and typical output results (e.g., regression coefficients, odds ratios, R^2, fixed effects vs. random effects, direct effects vs. indirect effects). Through this systematic comparison, the aim is to reveal the correspondence between model selection and data characteristics, providing methodological references for future research.

2. Literature Review

2.1 Logistic Regression

Logistic regression is applied when the dependent variable is binary. A binary response variable indicates an outcome with only two possible categories, such as survival versus death, or disease versus health. In practice, these two categories are usually coded as 1 and 0. Logistic regression estimates the probability of an event occurring through the logit function, expressed as:

$$\text{logit}(p) = a + b_1x_1 + b_2x_2 + ... + b_ix_i. \quad (1)$$

Here, p represents the probability of the outcome, the x terms are independent variables, and the b terms are the corresponding coefficients. Since it is often difficult to directly interpret the coefficients b in practical applications, they are usually converted into odds ratios (ORs). An OR indicates how many times the odds of the outcome change when a given independent variable increases by one unit, while other variables are held constant. The most important outputs of logistic regression are the ORs and the coefficients b, and additional statistics such as the standard error (SE), Wald χ^2, p-values, and 95% confidence intervals (CIs) are typically reported to assess the significance and reliability of the model [12].

In studies of adolescent depression, the dependent variable is usually whether depressive symptoms are present, while the independent variables are typically factors hypothesized to influence depression, such as gender and lifestyle.

In a large-scale national survey of Chinese children and adolescents, researchers applied logistic regression and OR values to identify factors such as gender, urban–rural residence, and physical activity level as significant contributors to depression risk [7].

Compared with multiple linear regression, logistic regression is more suitable for binary outcomes rather than continuous symptom scales.

2.2 Multiple Linear Regression (MLR)

A straight line is the fitted outcome of the Multiple Linear Regression (MLR) model, which is used to investigate the relationship between multiple independent variables and one dependent variable. With Y as the dependent variable, X terms as the independent variables, and b terms as the coefficients corresponding to each X, the function can be written as follows:

$$Y = a + b_1 \times X_1 + b_2 \times X_2 + ... + b_k \times X_k. \quad (2)$$

The MLR output's most crucial values are the b values. A coefficient's sign indicates whether the relationship with Y is positive or negative, while its absolute value shows how strongly the related independent variable influences Y. Additionally, MLR produces a value known as R2, which is a measure of the model's ability to explain the data. A model's appropriateness cannot be adequately described by a single R2 value; instead, it should be assessed using additional metrics like SE, t-statistic, p-value, and confidence intervals. Since strongly correlated variables can result in severe estimation bias, the stability of MLR is largely dependent on the independence of independent variables [8,11].

In a study on adolescent depression, MLR was utilized to examine the connection between the resilience of adolescents with depression and other characteristics, including family structure, childhood traumatic experiences, and adolescent weight. It was discovered that eight of the components significantly impacted the resilience of depressed teenagers by looking at the coefficients that corresponded to each factor and combining the SE, p-value, and t-value [8].

Multiple linear regression is better suited for continuous outcomes than logistic regression, but it is unable to account for hierarchical data structures or individual variability.

2.3 Linear Mixed-effects Model (LMM)

A linear mixed-effects model (LMM) is distinguished by its inclusion of both random and fixed variables, allowing it to account for individual differences and reflect the relationship between the variable of interest and the dependent variable better. Random variables refer to individual differences, such as students from different schools. Fixed variables are variables of interest that remain constant throughout the experiment. LMMs incorporate random intercepts and random slopes, allowing different individuals or groups to have their own baseline levels and effect sizes within the model [14].

In a study examining the relationship between parenting methods and the mental health of parents and children, researchers collected data on multiple families at different time points. They identified factors such as Structured Parenting, Shared Parenting, household income, parental race, and child gender as fixed effects, while treating each family and participant as random effects. The output included coefficients and p-values for each independent variable, revealing a significant association with Shared Parenting, with parents and children using this approach exhibiting lower rates of depression [6].

Unlike simple logistic or linear regression, mixed-effects models explicitly account for within-subject correlations and hierarchical structures, though they are more complex to specify and interpret.

2.4 Mediation Model

The mediation model can be used to analyze how an independent variable affects a dependent variable by introducing a mediator variable. There can be multiple mediator variables, and the principle of the mediation model is briefly described using one mediator variable as an example. In the mediation model, regression equations are used to describe the relationships between variables:

$$Y = cX + e_1,\ M = aX + e_2,\ Y = c'X + bM + e_3. \quad (3)$$

Here, X serves as the predictor, Y represents the outcome, and M functions as the mediator. In this specification, the coefficient a quantifies the influence of X on the mediator M, while b captures the effect of M on the outcome Y. The path coefficient ccc represents the total association between X and Y, whereas c' reflects the remaining direct effect of X on Y once the mediating role of M has been taken into account. The mediation effect is tested by sequentially examining whether the regression coefficient c and the product ab are significant. If both are significant, the mediation effect is considered significant [13].

In research on adolescent depression, when researchers identify a factor that has a significant relationship with depression, they can use a mediation model to study how this factor influences the development of depression in adolescents. Yu X. et al. hypothesized that thwarted belongingness plays an important role in the relationship between social exclusion and depression [9]. They established a mediation model with thwarted belongingness as the mediator variable, social exclusion as the independent

ISSN 2959-6157

variable, and depression as the dependent variable. The study found that thwarted belongingness partially explains the link between social exclusion and depression.

Unlike logistic or linear regression, mediation models focus less on prediction and more on explaining pathways between risk factors and outcomes.

In summary, data types and methodological approaches in adolescent depression research vary considerably. Multiple linear regression and logistic regression are typically applied to cross-sectional data and produce results that are relatively easy to interpret. In contrast, linear mixed-effects models are more suitable for longitudinal data, while mediation models can be applied to both longitudinal and cross-sectional designs, although the interpretation of the latter two methods is generally more complex.

3. Methodology

To ensure the review reflects the latest research on adolescent depression, the literature search primarily focused on studies published within the past five years (2020–2025). However, when introducing the theoretical foundations of statistical models, some earlier classic references were included to ensure the authority of the conceptual explanations.

The literature search primarily utilized the PubMed, Web of Science, PsycINFO, and Embase databases. Search terms employed included “adolescent depression”, “adolescent mental health”, “risk factors”, and “statistical models” combined with “linear mixed effects”, “multivariate linear regression”, “logistic regression”, and “multivariate regression”. These search terms were used to ensure high relevance between the literature and the objectives of this review. Furthermore, only peer-reviewed articles were considered to guarantee the rigor and reliability of the literature.

Subsequent exclusion criteria were applied to further refine the most suitable literature. Articles were excluded if: the primary focus population was not adolescents; the primary research subject was a disorder other than depression; the research methodology, data sources, and conclusions were not clearly stated; or none of the four targeted models were employed. Following this screening, the retained articles primarily focused on studies best illustrating the application of these four statistical models in adolescent depression research.

For each selected study, details were systematically extracted regarding data collection methods, including whether questionnaires, interviews, or longitudinal follow-ups were used. The sampling strategy of each article was recorded, such as random sampling, stratified sampling, or cohort recruitment. The type of data was identified, distinguishing between cross-sectional datasets, longitudinal observations, or studies with multiple outcome measures. The specific statistical model employed in each study was noted together with the research problem it was applied to address, such as predicting depression scores, identifying demographic and lifestyle risk factors, or modeling changes in symptoms over time. In addition, the form of the reported results was documented, including regression coefficients, odds ratios, or variance estimates, along with the conclusions that the authors directly drew from these outputs.

4. Results

Table 1 presents a comparison of four statistical models that have been frequently applied in studies on the etiology of adolescent depression: linear mixed-effects models, multiple linear regression, logistic regression, and the mediation model. The models differ in terms of their dependent and independent variable requirements, interpretation of results, appropriate research scenarios, and data type.

Table 1. Comparison of Statistical Models in Adolescent Depression Etiology Research

Comparison Dimension	Logistic Regression	Multiple Linear Regression	Linear Mixed-Effects Model	Mediation Model
Independent variables	Continuous, categorical (dummy coded)	Continuous, categorical (dummy coded)	Fixed: continuous; Random: categorical (dummy coded)	Continuous, categorical (dummy coded)
Dependent variable	Binary	Continuous	Continuous (within-cluster correlation allowed)	Continuous / binary
Outputs	β(coefficient), OR, SE, Wald χ^2, 95% CI, p	β, R^2, p	β, CI, p, AIC, BIC, ICC	direct effect, indirect effect, total effect, significance tests of mediation

Readability	Medium (OR interpretation)	High (R^2 intuitive)	Medium (random effects complex)	Medium (direct/indirect decomposition)
Interpretation	Effect on odds of outcome	Mean change in Y per unit X	Fixed effects as in regression; random effects capture variance	Direct effect of X on Y; indirect effect via M
Use cases	The effect of family cohesion on the likelihood of adolescents reporting depressive tendencies (binary outcome: presence vs. absence of depression)	The effect of social support and depression level on adolescents' CD-RISC resilience scores. (continuous outcome: total scale score)	The longitudinal effect of parenting style (structured vs. shared) on children's depressive symptoms across different developmental stages.	The effect of social exclusion on adolescent depression mediated by sense of belonging.
Data type	Cross-sectional	Cross-sectional	Longitudinal / hierarchical	Cross-sectional / longitudinal

Logistic regression and multiple linear regression were primarily applied to cross-sectional data. When used with longitudinal data, they required additional procedures to account for within-subject correlations or otherwise necessitated the use of more advanced models. Cross-sectional data refer to samples collected at a single time point. In contrast, linear mixed-effects models were mainly employed to analyze longitudinal or hierarchical data, such as repeated measurements on the same individual across different time points or students nested within classes and schools. Mediation models could be applied to both cross-sectional and longitudinal data, depending on the type of regression specified for each pathway.

Across all four models, independent variables could be either categorical or quantitative, although categorical predictors needed to be numerically coded before inclusion. Multiple linear regression and linear mixed-effects models were used to analyze continuous dependent variables, while logistic regression was typically applied to binary outcomes. Mediation models were flexible, accommodating either continuous or binary dependent variables.

Differences were also observed in the mathematical formulation and reported outputs of the four models. Logistic regression applied the logit link function to model binary outcomes, whereas multiple linear regression expressed continuous outcomes as a linear combination of predictors. Linear mixed-effects models incorporated both fixed and random effects to address longitudinal or hierarchical continuous outcomes. Mediation models employed multiple regression equations simultaneously, decomposing the overall association into direct and indirect pathways.

In terms of reported outputs, logistic regression studies typically presented regression coefficients, odds ratios, 95% confidence intervals, and significance tests. Multiple linear regression studies emphasized regression coefficients and R^2 values as indicators of explained variance. Linear mixed-effects models produced both fixed-effect estimates and random-effect variance components, often accompanied by model fit indices such as AIC, BIC, or ICC. Mediation models reported direct, indirect, and total effects, with significance most frequently assessed by bootstrap confidence intervals or Sobel tests.

5. Discussion

In the comparison of four models, logistic regression and multiple linear regression are relatively straightforward approaches. They are simpler in terms of both data requirements and model specification, and their results are comparatively easier to interpret. Logistic regression is typically applied when the dependent variable is binary, which is common in adolescent depression research, for example, "whether the individual has depression" or "whether comorbid conditions are present." Because data in this field are primarily derived from questionnaires, the design and scoring of survey items often directly determine the choice of statistical model. When items are presented in a "yes/no" format, logistic regression is generally selected, whereas for items with continuous responses, such as rating distress severity on a scale from 1 to 5, multiple linear regression is more appropriate.

The advantage of multiple linear regression lies in its straightforward interpretability: researchers can examine the slope and significance of each predictor to determine whether it constitutes a significant risk factor, and the sign of the coefficient indicates whether the effect is positive or negative. This simplicity explains the wide use of logistic and linear regression in adolescent depression studies. However, in practice, the data collected are often more complex. Even within the same school, students differ considerably in personality and physical health, creating heterogeneity that simple regression models cannot adequately capture. For example, when examining the association between sleep quality and adolescent depression,

ISSN 2959-6157

students differ in physical condition, family environment, and academic stress. If such individual-level differences are not accounted for, the results may misleadingly attribute the effect to sleep quality when it is in fact driven by underlying heterogeneity.

Multiple linear regression does not fully account for such individual differences. It averages these differences into the dependent variable, which may result in substantial model bias. In such cases, an observed effect of a predictor might be driven more by individual variability than by the predictor itself. Linear mixed-effects models address this limitation by introducing random intercepts and random slopes into the equation, thereby incorporating individual-level variation into the modeling process. This allows for more accurate estimation of the true effect of predictors on the outcome. Beyond accounting for individual differences, these models are also well-suited for longitudinal data analysis. Longitudinal designs are an important approach in adolescent depression research, involving repeated measurements of the same participants over extended periods to examine how specific factors exert influence across time. Consequently, when researchers intend to conduct long-term follow-up studies of individuals or groups, linear mixed-effects models represent an appropriate choice.

Compared with the other three models, the mediation model occupies a distinct position. Before applying this model, researchers generally need to establish through the preceding models or other methods researchers must initially establish, often via preliminary regression models, that a statistically reliable association exists between the predictor and the response variable. Once this is confirmed, potential mediators can be identified based on existing theory and empirical findings, and mediation models can then be used to test whether these mediators significantly account for the observed relationship.

Overall, in adolescent depression research, the choice of model largely depends on the method of data collection and the research question under investigation. If the aim is simply to identify risk factors significantly associated with depression, multiple linear regression, logistic regression, or linear mixed-effects models may be selected depending on data type and measurement. However, if the objective is to uncover the mechanisms or intermediate pathways through which independent variables influence outcomes—that is, to answer "how" or "why" such effects occur—mediation models become a valuable option.

Future research should also consider integrating and expanding these approaches. Although logistic regression, multiple linear regression, mixed-effects models, and mediation models each have their appropriate applications, none of them can fully capture the multifaceted etiology of adolescent depression. Future studies may therefore need to adopt more complex models or combine multiple methods. For instance, classical regression techniques could be integrated with machine learning approaches such as random forests or support vector machines to handle high-dimensional predictors, while regression frameworks remain valuable for ensuring interpretability. In terms of research design, many current studies on adolescent depression rely heavily on self-reported questionnaires, which are prone to bias, as some participants may not accurately report their condition. Future work should incorporate more diverse forms of data, such as clinical assessments and objective measurements obtained from instruments. Finally, greater emphasis should be placed on cross-cultural and interdisciplinary applications, as the risk factors and mechanisms of depression may vary substantially across different populations and environments.

6. Conclusion

The prevalence of adolescent depression has increased substantially, and once developed, it leads to multiple adverse consequences that seriously affect adolescents' lives. Investigating the etiology of adolescent depression and implementing early interventions are therefore of great importance. A decisive factor in ensuring the reliability of research findings is whether appropriate statistical methods are employed to analyze the data. This review compared four widely used statistical approaches in adolescent depression research: logistic regression, multiple linear regression, linear mixed-effects models, and mediation models. The findings indicate that the choice of statistical model largely depends on how data collection is designed: logistic regression is suitable for binary outcomes such as "presence or absence of depression"; multiple linear regression is applied to continuous outcomes such as symptom scores; linear mixed-effects models are suitable for analyzing longitudinal data following a cohort over time; and mediation models can reveal how an already identified risk factor influences the development of depression through specific mechanisms.

However, most current studies still rely heavily on self-reported questionnaires, which are prone to bias, and the integration of different models remains limited. Future research should explore the combination of multiple data sources, like clinical assessments and objective measurements, strengthen cross-disciplinary and cross-cultural comparisons, and adopt more advanced or hybrid modeling strategies in order to more comprehensively uncover the etiology of adolescent depression.

References

[1] Shorey Shefaly, Ng Esperanza Debby, Wong Celine H J. Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis. British Journal of Clinical Psychology, 2022, 61(3): 287–305.

[2] Beck A, LeBlanc J C, Morissette K, Hamel C, Skidmore B, Colquhoun H, Stevens A. Screening for depression in children and adolescents: A protocol for a systematic review update. Systematic Reviews, 2021, 10(1): 24.

[3] Iannattone S, Mezzalira S, Bottesi G, Gatta M, Miscioscia M. Emotion dysregulation and psychopathological symptoms in non-clinical adolescents: The mediating role of boredom and social media use. Journal of Affective Disorders, 2024, 356: 91–100.

[4] Kaywan P, Ahmed K, Ibaida A, Miao Y, Gu B. Early detection of depression using a conversational AI bot: A non-clinical trial. PLoS ONE, 2023, 18(1): e0279743.

[5] Zulfiker M S, Kabir N, Biswas A A, Nazneen T, Uddin M S. An in-depth analysis of machine learning approaches to predict depression. Current Research in Behavioral Sciences, 2021, 2: 100044.

[6] Cost KT, Mudiyanselage P, Unternaehrer E, et al. The role of parenting practices in parent and child mental health over time. BJPsych Open. 2023, 9(5): e147. doi:10.1192/bjo.2023.529

[7] Zhang Y, Li D, Li X, Wang Y. Demographic and lifestyle factors associated with depressive symptoms among children and adolescents in China. Chinese Journal of School Health, 2021, 42(5): 657–661.

[8] Wang M, Li Q. Path analysis of factors influencing resilience in adolescents with depression. Journal of Peking University (Health Sciences), 2021, 53(5): 809–817.

[9] Yu X, Du H, Li D, Sun P, Pi S. The influence of social exclusion on high school students' depression: A moderated mediation model. Psychology Research and Behavior Management, 2023, 16: 4725–4735.

[10] Pedersen G A, Lam C, Hoffmann M, Zajkowska Z, Walsh A, Kieling C, Mondelli V, Fisher H L, Gautam K, Kohrt B A. Psychological and contextual risk factors for first-onset depression among adolescents and young people around the globe: A systematic review and meta-analysis. Early Intervention in Psychiatry, 2023, 17(1): 5–20.

[11] Schneider A, Hommel G, Blettner M. Linear regression analysis: Part 14 of a series on evaluation of scientific publications. Deutsches Ärzteblatt International, 2010, 107(44): 776–782. https://doi.org/10.3238/arztebl.2010.0776

[12] Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. Critical Care, 2005, 9(1):112–118.

[13] Wen Z, Ye B. Analyses of mediation effects: The development of methods and models. Advances in Psychological Science, 2014, 22(5): 731–745.

[14] Brown V A. An introduction to linear mixed-effects modeling in R. Advances in Methods and Practices in Psychological Science, 2021, 4(1).