# 基于回归分析与随机森林的初中数学成绩影响因素分析

李冰
天津市梧桐中学，天津 300200
DOI:10.61369/ASDS.2026010004

**摘　　要：** 为探究影响八年级初中生数学成绩的关键因素，基于回归分析模型与随机森林方法研究了某中学八年级84名学生期末数学成绩的影响因素，系统分析了各解释变量与成绩间的相关性及影响机制。结果表明：期中成绩是核心预测指标，月考成绩、两次测试成绩与期末成绩呈现强正相关关系，期末前作业完成质量呈中等正相关，出勤天数对期末成绩影响较弱，性别因素存在显著差异，班级因素无显著影响。研究结果可为初中数学个性化教学与学业提升策略制定提供实证依据。

**关 键 词：** 回归分析；随机森林；数学成绩；影响因素

## Determinants of Middle-School Mathematics Achievement: A Regression–Random-Forest Hybrid Analysis

Li Bing
Tianjin Wutong Middle School, Tianjin 300200

**Abstract:** To identify the key influencing factors of eighth-grade mathematics achievement, we analyzed math scores for 84 students using a regression model and random-forest method. Our study systematically examined the strength and mechanism of each predictor's association with performance. Midterm score emerged as the dominant predictor; monthly-quiz and two additional test scores displayed strong positive correlations with the final result, whereas pre-final homework quality showed a moderate positive link. Attendance exerted only a weak influence, and a significant gender gap was detected; class membership had no discernible effect. These findings furnish an empirical foundation for tailoring middle-school mathematics instruction and designing targeted academic-improvement strategies.

**Keywords:** regression analysis; random forest; mathematics score; influencing factors

## 引言

初中阶段的学习是夯实学生基础知识、培养综合素养的关键环节，数学是重要载体之一，它能培养学生的逻辑思维、抽象推理与问题解决能力，数学成绩的高低会直接影响学生的理科学习路径与综合素养发展[1]。学生在这个阶段需要完成知识跨越（从直观认知发展至逻辑证明），学业难度与认知要求都显著提升，数学成绩不仅反映出学生的知识掌握水平，也能折射出学生的学习态度、习惯等深层素养。在初中数学教学过程中，学生成绩出现分化的情况非常普遍，而成绩差异受多重因素交织影响。

国内外学者对学业成绩影响因素的研究已形成丰富成果[2]，初中生数学学业成绩受学习行为[3]、个体特征层面、教学因素、个体心理与情绪[4]等多维度因素交互作用，其中阶段性练习（月考、期中）、出勤次数、平时作业效果也是关键作用变量。在个体特征层面，部分研究认为性别差异会影响数学学习，男性在空间推理、逻辑运算上表现更优，女性在细节处理、基础运算上更具优势[5]，也有研究指出，随着教育公平推进，性别对数学成绩的影响逐渐减弱[6]。在学习行为层面，出勤作为课堂参与的基础，其规律性与学业成绩呈正相关，缺课会导致知识断层，显著降低成绩表现[7]；作业作为课堂知识的延伸，其完成质量直接反映知识掌握程度，高质量作业能强化知识内化，提升学业成效[8]。Smith等研究发现阶段性练习（月考、期中）是学业评价的重要环节，月考成绩能及时反馈学生阶段性知识漏洞，其与期末成绩的相关系数达0.72，是重要预测指标[9]。王颖对初中学生的研究表明，期中成绩因覆盖知识范围与期末高度契合，相关性最强，可作为期末成绩的核心预测变量[10]。

现有研究虽已关注多因素对数学成绩的影响，但存在三方面不足：一已有研究多聚焦单一维度，或对影响因素的分析较为零散，缺乏对性别、出勤、作业完成度、阶段性练习等多指标的系统整合研究；二是针对八年级关键学段的实证研究较少；三是部分研究采用传统统计方法，对非线性关系的捕捉能力较弱。因此本文结合多种量化分析方法研究期末成绩与性别、出勤率、平时练习、期中成绩、作业得分等因素的相关性，为教学实践提供更精准的参考。

作者简介：李冰，女，天津市梧桐中学，二级教师，主要从事初中数学教学与研究工作。

 

## 一、数据来源与统计方法

### （一）数据来源

研究选取了八年级A班（46人）和B班（38人），共84名学生为研究对象，学生年龄分布为13-14岁，均接受统一数学课程教学，排除特殊教育需求学生。响应变量选取了学年期末数学成绩，解释变量考虑了性别、月考成绩、期中成绩、平时测试成绩、出勤天数及平时作业得分等，变量具体定义见表1。

表1：变量定义

| 变量类型 | | 变量含义 | 定义与测量方式 |
|---|---|---|---|
| 响应变量 | Y | 期末成绩 | 期末数学质量检测成绩 |
| 解释变量 | X1 | 班级 | 分类变量（A班、B班） |
| | X2 | 性别 | 分类变量（1=男生，2=女生） |
| | X3 | 月考成绩 | 月考数学成绩（满分100分） |
| | X4 | 期中成绩 | 上学期期中数学检测成绩 |
| | X51、X52、X53 | 测试成绩 | 三阶段的测试成绩 |
| | X61、X62、X63 | 出勤天数 | 三阶段的实际出勤天数 |
| | X71、X72、X73 | 作业得分 | 三阶段的数学作业得分（满分3分） |

注：如无特殊说明，成绩满分是100分。

### （二）统计分析方法

为系统探究各解释变量对期末数学成绩的影响方向与程度，揭示变量间的潜在关联规律。本研究整合多维度数据分析方法，构建由表及里、逐层递进的分析框架，从数据特征刻画、关联强度探究到模型构建与优化，形成完整的分析链路，具体方法如下：

描述性统计分析：测算各核心变量的均值、标准差、中位数、极值及偏度等关键统计量，通过定量指标精准刻画数据的整体分布特征、离散趋势与形态特征，为后续深入分析奠定数据基础。

相关性分析：剖析各自变量与期末数学成绩的线性关联强度，构建相关性矩阵以可视化呈现变量间的关联模式，初步筛选对期末成绩存在潜在影响的关键变量。

多元线性回归：构建递进式回归模型，通过引入不同层次的自变量，结合假设检验方法，依次验证各因素对期末成绩影响的统计学显著性，明确变量间的线性作用关系。

随机森林模型：基于集成学习算法构建随机森林模型，评估各输入变量的特征重要性排序，有效弥补传统线性回归模型难以捕捉变量间复杂非线性关系的局限，挖掘影响期末成绩的核心驱动因素。

LASSO回归：依托L1正则化机制，对多元线性回归模型进行优化，通过系数压缩与变量筛选，剔除冗余变量，降低模型过拟合风险，进一步提升模型的泛化能力与预测效能。

## 二、数据分布特征

### （一）描述性统计

变量的详细描述性统计分析见表2。

表2：变量的描述性统计分析

| 变量名 | 变量含义 | 均值 | 标准差 | 中位数 | 最小值 | 最大值 | 偏度 |
|---|---|---|---|---|---|---|---|
| Y | 期末成绩 | 56.10 | 25.10 | 58.00 | 10.0 | 100 | −0.07 |
| X3 | 月考成绩 | 67.95 | 22.97 | 72.00 | 17.0 | 100 | −0.42 |
| X4 | 期中成绩 | 76.30 | 18.60 | 81.50 | 21.0 | 100 | −0.97 |
| X51 | 测试成绩1 | 67.52 | 23.83 | 72.00 | 0.0 | 100 | −0.55 |
| X52 | 测试成绩2 | 68.10 | 28.28 | 76.00 | 0.0 | 99 | −1.31 |
| X53 | 测试成绩3 | 52.79 | 27.50 | 51.00 | 0.0 | 100 | −0.18 |
| X61 | 出勤天数1 | 15.39 | 4.98 | 12.00 | 6.5 | 21 | 0.08 |
| X62 | 出勤天数2 | 17.68 | 1.59 | 18.00 | 10.0 | 19 | −1.84 |
| X63 | 出勤天数3 | 17.58 | 3.55 | 17.00 | 1.5 | 21 | −1.53 |
| X71 | 作业得分1 | 1.91 | 0.75 | 2.00 | 0.0 | 3 | −0.65 |
| X72 | 作业得分2 | 1.99 | 0.90 | 2.20 | 0.0 | 3 | −0.83 |
| X73 | 作业得分3 | 1.79 | 0.93 | 2.09 | 0.0 | 3 | −0.65 |

由表2可知，响应变量期末成绩（Y）的均值56.10分，低于60分及格线，标准差25.10，均值较低且离散程度大，反映学生成绩分化明显. 期末成绩主要集中在40-60分区间（约25人），呈近似对称分布，偏度为-0.07，表明近似呈现正态分布。解释变量中期中成绩（X4）均值为76.30分，高于月考成绩（67.95分），说明学生中期学习效果提升。三阶段的出勤天数均值分别为15.39、17.68、17.58天，相差不太大，但第二次出勤天数最高，其偏度-1.84，呈左偏分布，说明多数学生二阶段的出勤天数较高，少数学生缺勤较多. 三次作业得分均值在1.79-1.99之间，标准差较小（0.75-0.93），说明学生作业完成质量差异不大。

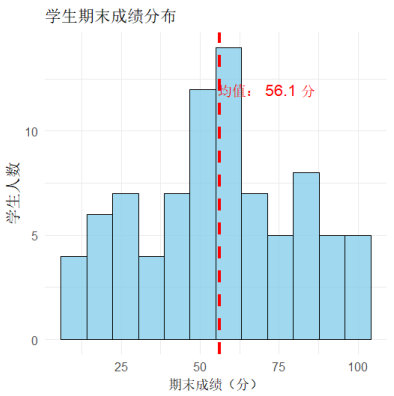


图1：学生期末成绩直方图

 

为了清晰展示学生期末数学成绩的分布特征，绘制了响应变量的直方图，见图1。由图1可知，期末数学成绩主要集中在25至75分之间，该区间人数最多，75至100分区间人数较少。整体成绩偏低，均值仅为56.1分，表明多数学生成绩未达到及格水平，成绩分布呈现右偏形态，高分段学生极少，整体学习效果可能不理想，需关注学生学习中重要的影响因素。

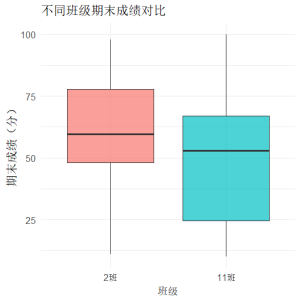


图2：两个班成绩的箱线

两个班的成绩对比箱线图如图2所示。由图2可得，两个班成绩对比来看，A班（红色）成绩中位数高于B班（绿色），且A班箱子厚度比B班小，说明A班成绩较为集中，A班整体成绩更优，B班成绩分化较严重。

**（二）相关性分析**

由于期末成绩与期中成绩具有较高的相关性，因此绘制了两个班成绩的散点图及拟合曲线，见图3。

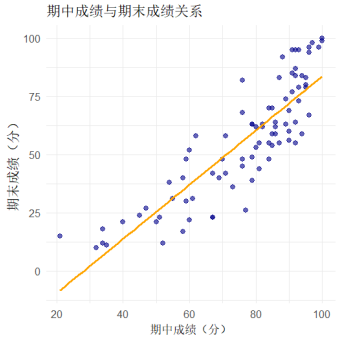


图3：期中成绩与期末成绩关系散点图

图3显示趋势线向上，表明期中成绩与期末成绩呈现强正相关关系。期中成绩高于80分的学生，期末成绩多大于70分，反之，期中成绩低于50分的学生，期末成绩多小于 50 分，表明期中成绩是期末成绩核心预测指标。

其他变量的相关性矩阵热力图见图4。

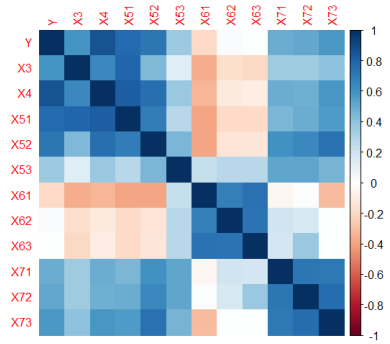


图4：变量的相关性矩阵热力图

第一次测试成绩（X51）相关性0.79，第二次测试成绩（X52）相关性0.75，均为强正相关，第三次作业得分（X73）相关性0.62，为中等正相关，期中成绩、测试成绩、平时作业得分与期末分数的相关性说明了数学成绩是一个具有连续性、可积累的结果，日常学习的过程性表现能直接反映并影响最终的学业水平。出勤天数（X61-X63）与期末成绩相关性均小于0.2，影响较弱，说明数学分数的呈现比较依赖基础知识是否牢固，不受偶尔缺勤的影响，也存在人虽然到了课堂，但可能上课走神、不参与课堂互动、不跟着老师的思路思考问题，这种 “无效出勤” 也无法转化为成绩的提升，也会呈现出勤天数和成绩之间没有明显关联的结果。

## 三、回归与随机森林模型分析

**（一）多元线性回归分析**

为了精准探究各解释变量对学生的影响关系，通过多元线性回归模型进行分析，结果见表3。

表3：基于多元线性回归的成绩的影响因素分析

| 估计 | 模型1 | 模型2 | 模型3 |
|---|---|---|---|
| 截距 | −32.83*** | −50.22*** | −36.25*** |
| 水平一层变量 | | | |
| X3(月考成绩) | | 0.11 | 0.12 |
| X4(期中成绩) | 1.17*** | 0.94*** | 0.92*** |
| test_mean(综合测试) | | 0.19 | 0.23* |
| attend_mean(综合出勤) | | 0.19 | |
| 水平二层变量 | | | |
| 班级 (11班) | | | 5.78 |
| 性别 (2) | | | −5.39* |
| $R^2$ | 0.746 | 0.773 | 0.782 |

注：*、**、*** 分别表示显著性水平为0.05、0.01、0.001。

由表3可得如下结论：

模型1仅含相关系数最大的期中成绩，$R^2$ 为0.746，说明期中成绩单独可解释期末成绩 74.6% 的变异，期中成绩每提高1分，期末成绩预期提高1.17分。在模型1的基础上，在加入水平一层的解释变量，构成了模型2的结果，$R^2$ 有所提升，达到0.773，期中成绩系数降至0.94（p<0.001），仍高度显著，表明期中成绩仍然是解释期末成绩的关键影响因素。月考成绩、综合测试、综合出勤的系数均不显著，说明这些变量在控制期中成绩后，独立贡献有限 . 在模型2基础上再加入班级、性别等因素构成模型3，$R^2$ 进

 

一步提升至0.782，期中成绩系数稳定在 0.92（p<0.001），综合测试成绩变得显著，表明综测成绩每提升1分，期末成绩能有效增加0.23分。性别因素显著，相较于男同学，女同学的成绩低5.39分。班级因素无显著差异。

### （二）随机森林方法

为避免样本划分的随机性，让评估结果更稳定可靠。通过模拟生成了不同样本划分下的超参数选择，以 RMSE（均方根误差）、$R^2$（判定系数）与 MAE（平均绝对误差）作为衡量预测模型精度的核心指标为评估指标。RMSE、MAE 值越小代表模型越好，$R^2$则越大越好。通过5折重复交叉验证，对12种超参数组合进行筛选，最优超参数选择见表4。

表4：随机森林超参数选择

| mtry | splitrule | RMSE | | MAE |
|---|---|---|---|---|
| 2 | variance | 13.63175 | 0.7303722 | 11.34675 |
| 2 | extratrees | 14.57091 | 0.7109171 | 11.87205 |
| 4 | variance | 13.00682 | 0.7441912 | 10.64079 |
| 4 | extratrees | 13.64631 | 0.7349649 | 11.09301 |
| 6 | variance | 12.85370 | 0.7475450 | 10.37788 |
| 6 | extratrees | 13.34373 | 0.7427058 | 10.90067 |
| 8 | variance | 12.77588 | 0.7490694 | 10.26280 |
| 8 | extratrees | 13.21776 | 0.7436431 | 10.75150 |
| 10 | variance | 12.79958 | 0.7453868 | 10.28632 |
| 10 | extratrees | 13.06458 | 0.7495955 | 10.64378 |
| 13 | variance | 12.85875 | 0.7422460 | 10.36590 |
| 13 | extratrees | 12.97155 | 0.7529319 | 10.57561 |

由表4可知，特征随机选择数为8，叶子节点最小样本数为5，决策树数量选择了800棵。在测试集和训练集学生成绩数据进行评估，评估结果见表5。

表5：模型性能评估

| 数据集 | 样本量 | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| 训练集 | 68 | 5.535 | 4.408 | 0.950 |
| 测试集 | 16 | 10.862 | 9.707 | 0.820 |

由表5，$R^2$ 越接近 1，说明模型能解释的因变量（期末成绩）变异比例越高，解释力越强 . 模型 $R^2$ =0.82说明在期末成绩的变化中有82% 可以由模型的自变量（比如作业得分、期中成绩、测试分数等）解释，剩下变化影响可能来自模型未纳入的因素（比如考试心态、临场发挥等），模型泛化性能良好，未出现明显过拟合。

为了更清晰展示解释变量的重要性，绘制了解释变量重要性程度图，见图5。

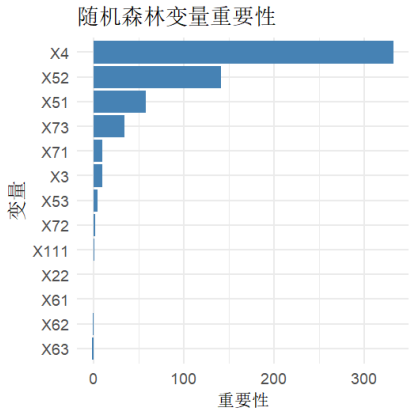


图5：随机森林变量重要性排序

由图5可知，期中成绩（X4）为最重要变量，对期末成绩影响最大，说明期中是阶段知识的综合验收，其覆盖范围、考查深度、评估维度最接近期末，且直接反映学生对中期核心知识的掌握程度，而这部分知识是期末备考的基础；X52、X51（测试成绩）次之；作业得分3（X73）排名第4，说明期末前的作业完成质量对成绩提升有针对性作用；出勤天数变量（X61–X63）重要性较低，班级与性别影响可忽略，表明成绩差异主要来自学生个体学习表现。

### （三）LASSO 回归

为避免模型过于简单而欠拟合，又防止模型过度复杂而出现过拟合，使用 LASSO 回归选择解释变量，通过交叉验证确定两个关键参数，最小化交叉验证误差 λ.min=0.206以及最小误差一个标准差内的最简模型 λ.lse=2.108。变量筛选结果见表6。

表6：变量筛选结果（λ.lse 模型）

| 变量 | 系数 |
|---|---|
| (Intercept) | −41.816 |
| 性别 X2 | −1.99 |
| 月考成绩 X3 | 0.006 |
| 期中成绩 X4 | 0.796 |
| 测试1成绩 X51 | 0.154 |
| 测试2成绩 X52 | 0.080 |
| 出勤 X62 | 1.221 |

由 LASSO 变量筛选结果可知，最重要的解释变量为性别、月考成绩、期中成绩、测试1成绩、测试2成绩与出勤 X62。期中成绩每提高1分，期末成绩预计提高0.796分。出勤天数每增加1天，期末成绩预计提高1.221分。两次测试成绩对期末成绩均有正向影响，性别差异也会对期末成绩产生影响。

## 四、结论与启示

### （一）研究结论

考虑数学知识具有层层递进、环环相扣的特点，比如 “全等三角形” 是 “轴对称” 的基础，“一次函数” 又会关联 “反比例函数”，那么月考、期中这类阶段性检测内容是覆盖了一个阶段的模块知识，分数高、成绩好则意味着学生已经梳理明白了这一模块的知识逻辑，掌握了扎实的基础知识；而反过来，分数较低，成绩偏低的学生并未掌握知识，知识存在漏洞，如果不及时补上，会影响后续新知识的学习，最终体现在期末成绩上。相比之下，作业、练习更偏向 “单点知识巩固”，对知识体系的检验力度远不如阶段性检测作业完成质量对期末成绩的积极作用。因此如研究结果显示期末成绩的最有效的预测指标是期中成绩，期中成绩与期末成绩相关性最强，月考成绩和两次阶段性测试成绩也呈强正相关。

出勤天数对期末成绩影响较弱，因为出勤”只是学习的 “必要非充分条件”，它无法反映知识掌握程度，是出勤过程中有效的学习行为在影响成绩，因此出勤天数只是一个 “表面指标”，和成绩的关联自然很弱。

性别因素存在显著差异，“显著” 不是 “差距大”，是

 

指这种差异并非由随机抽样的误差而导致的，是真实存在的规律，因为不同性别在数学学习上的思维特点、学习习惯会有所不同（比如男生擅长逻辑推理类题型，女生擅长计算和规范答题类题型）。

班级因素无显著影响，说明成绩差异主要源于学生个体学习表现，班级教学环境差异并不会引起显著影响，而是考虑期末成绩的核心影响因素的解释力很强（例如阶段性检测成绩、作业得分等），当模型中纳入了这些核心变量后，班级因素的影响就被“稀释”了——学生的成绩好坏，主要由他自身的阶段性表现决定，与学生所在的哪个班关系不大。

成绩分化的话，B 班较 A 班更严重，说明 B 班学生在月考、期中这类练习后的补漏效果分层明显，基础扎实的学生能及时查漏补缺巩固知识，基础薄弱的学生知识漏洞本来就多，因此漏洞越积越多，也就导致差距逐渐拉大，自律的学生能通过作业和检测持续提分，而基础弱、学习习惯差的学生则跟不上，叠加之下会进一步加剧整体分化。

**（二）教学启示**

从研究结果来看，提示教学时重视强化阶段性检测的反馈：重视期中、月考等阶段性测试的诊断功能，针对测试中暴露的知识漏洞及时开展专项辅导，尤其关注期中成绩低于50分的学生，提前干预避免成绩进一步下滑。

优化作业设计与反馈：提高作业针对性，尤其加强期末前的知识巩固类作业，通过高质量作业强化知识内化；同时关注作业完成质量差异较小的特点，设计分层作业满足不同学生需求。

实施个性化教学：针对 B 班等成绩分化严重的班级，建立分层教学机制，对学困生加强基础辅导，对优等生提供拓展性学习任务，缩小成绩差距。重视性别因素带来的成绩差异，因材施教，为不同性别学生提供适配的学习指导策略，同时避免因班级标签忽视个体发展需求。

## *参考文献*

[1] 教育部 . 义务教育数学课程标准（2022 年版）[S]. 北京：北京师范大学出版社 ,2022.

[2] 李娟 . 初中数学学业成绩影响因素的实证研究 [J]. 教育学报 ,2020,16（3）:89-96.

[3] 刘伟 . 项目式教学对八年级数学成绩的影响研究 [J]. 课程 · 教材 · 教法 ,2021,41（5）:112-117.

[4] 张明 . 学习兴趣对初中生数学成绩的影响及培养策略 [J]. 数学教育学报 ,2019,28（2）:45-50.

[5]Hyde J S. Gender Differences in Mathematics Performance:A Meta-Analysis [J]. Psychological Bulletin,2005,131（2）:109-135.

[6] 王静 . 教育公平视角下初中生数学成绩的性别差异研究 [J]. 教育研究 ,2022,43(7）:68-75.

[7] 陈阳 . 初中生课堂出勤与学业成绩的相关性分析 [J]. 基础教育参考 ,2020（12）:34-37.

[8] 刘敏 . 数学作业质量对初中生学业成绩的影响机制研究 [J]. 数学通报 ,2021,60(8）:23-28.

[9]Smith A,Jones B. The Role of Monthly Tests in Predicting Final Mathematics Achievement of Middle School Students [J]. Journal of Educational Measurement,2020,57（3）:412-428.

[10] 王颖 . 初中生期中与期末数学成绩的相关性研究 [J]. 中国教育学刊 ,2022(S1）:156-158.

 