《信息分析》课程作业

专业： 信息管理学院 **学号：** 211820073 **姓名：** 胡涂

学生成绩影响因素研究

摘要：当今社会，教育的重要性日益凸显，学生成绩也逐渐成为了评价教育质量的重要指标之一。因此，了解学生成绩的影响因素，对于改善教育质量，提高学生学习效果具有重要意义。本研究通过收集大量学生的个人信息、家庭背景信息等方面的数据，利用 R 语言，采用多元线性回归的方法分析学生成绩与各因素之间的关系。结果表明，影响学生成绩的关键因素包括：家庭经济状况，家长教育背景和受教育程度，学生自身学习能力和学习习惯等。这些结果为学生成绩提升以及提高教育教学质量提供了一定的理论依据和实践指导，对于学校和决策者在进行教育规划和教学改进方面具有一定的参考价值。

关键词：教育；多元线性回归；特征选择

一、 引言

近年来，教育的重要性和学生成绩的评价机制逐渐成为全球关注的焦点。学生成绩的优劣不仅可能影响到学生的升学、就业和人生的方向选择，同时也反映了学校和国家的教育教学水平。因此，提高学生成绩是教育教学改革和发展的重要目标之一，而了解学生成绩的影响因素与学习效果的关联，则是提高学生成绩的关键所在。

虽然学生的学习成绩受许多因素的影响，部分因素是个人内部因素，如天赋、个人性格等，但更多的因素是来自环境和社交方面的，包括家庭背景、学校教育水平、教学质量等。因此，许多学者和研究者对影响学生成绩的因素进行了深入研究和探讨，试图找出其中的相关关系，为提高学校教学质量和学生成绩提供理论和应用支持。

本研究旨在探讨影响学生成绩的关键因素，以及这些因素之间的相互关系。本文基于已有的文献和数据，借助统计学方法和机器学习技术，对影响学生成绩的因素进行研究，寻找其中的相关性和规律性。本研究建立了多元线性回归模型，预测了学生的学习成绩，并在此基础上提出了改善学生成绩的建议和措施，为学校和决策者在进行教育规划和教学改进方面提供参考和支持。

本研究的意义与价值在于：一方面，深入探讨影响学生成绩的因素，为教育教学提供理论支持和实践指导；另一方面，本研究的研究方法和实验方案具有较大的参考价值，可为类似领域的研究提供新的思路和视角。

二、 学生成绩影响因素的量化评估

（一）、 数据来源与介绍

数据来源自 Kaggle 平台(https://www.kaggle.com/datasets/dipam7/student-grade-prediction)。

该数据接近两所葡萄牙学校中学教育的学生成绩。数据属性包括学生成绩、人口统计、社会和学校相关特征），它是使用学校报告和问卷收集的。提供了两个关于两个不同学科表现的数据集：数学和葡萄牙语。

数据集中共有 395 条记录，共 33 个变量，数据字典如下。

字段	含义
school	学生所在的学校（二元属性：'GP'-Gabriel Pereira 或'MS'-Mousinho da Silveira）
sex	学生的性别（二元属性：'F'-女性或'M'-男性）
age	学生的年龄（数值属性：从 15 到 22 岁）
address	学生的家庭住址类型（二元属性：'U'-城市或'R'-农村）
famsize	家庭规模（二元属性：'LE3' - 小于或等于 3 人或'GT3' - 大于 3 人）
Pstatus	父母的同居状况（二元属性：'T'-共同居住或'A'-分居）
Medu	母亲的教育程度（数值属性：0-无，1-小学教育(4 年级)，2-5 年级至 9 年级，3-中等教育或 4-高等教育）
Fedu	父亲的教育程度（数值属性：0-无，1-小学教育(4 年级)，2-5 年级至 9 年级，3-中等教育或 4-高等教育）
Mjob	母亲的职业（名义属性：'teacher'，'health' care related，'services'（如行政或警察），'at_home'或'other'）
Fjob	父亲的职业（名义属性：'teacher'，'health' care related，'services'（如行政或警察），'at_home'或'other'）
reason	选择这所学校的原因（名义属性：靠近'home'，学校 'reputation'，课程 'preference'或'other'）
guardian	学生的监护人（名义属性：'mother'，'father'或'other'）
traveltime	家到学校的路上时间（数值属性：1-＜15 分钟，2-15 到 30 分钟，3-30 分钟到 1 小时或 4-＞1 小时）
studytime	每周学习时间（数值属性：1-＜2 小时，2-2 到 5 小时，3-5 到 10 小时或 4-＞10 小时）
failures	过去的班级不及格次数（数值属性：如果 1≤n<3，则为 n，否则为 4）
schoolsup	额外的教育支持（二元属性：是或否）
famsup	家庭教育支持（二元属性：是或否）
paid	课程科目中的额外有偿课程（数学或葡萄牙语）（二元属性：是或否）
activities	课外活动（二元属性：是或否）
nursery	是否参加过托儿所（二元属性：是或否）

higher	想要继续接受高等教育（二元属性：是或否）
internet	家庭是否有互联网（二元属性：是或否）
romantic	是否有恋爱关系（二元属性：是或否）
famrel	家庭关系质量（数值属性：从 1-非常糟糕到 5-非常好）
freetime	放学后的空闲时间（数值属性：从 1-非常少到 5-非常多）
goout	和朋友出去玩（数值属性：从 1-非常少到 5-非常多）
Dalc	工作日饮酒量（数值属性：从 1-非常少到 5-非常多）
Walc	周末饮酒量（数值属性：从 1-非常少到 5-非常多）
health	当前健康状况（数值属性：从 1-非常差到 5-非常好）
absences	学校缺勤次数（数值属性：从 0 到 93）
G1	第一学期成绩
G2	第二学期成绩
G3	第三学期成绩

本研究的目的就是基于前 29 个变量（除去学校类别影响），逐步预测 G1、G2、G3 三个学期的成绩，分阶段考察出影响学生成绩的主要因素，并加以解释。

值得注意的是，在模型中，大多数因变量的影响并不显著，我们需要一些特征选择的方法来约简模型并处理异常值。本文采用了根据显著性提取特征与随机森林提取特征的方法，前者并不能达到最优的效果，而后者仅在预测 G3 时使用。

（二）、 对第一学期成绩的拟合与解释

拟合采用显著性（即系数是否为 0）的方法对模型进行约简，因此先对所有变量进行回归（除了 G2 与 G3），同时考虑了截距项是否为 0 的情况，发现截距为 0，模型拟合优度更高。

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
sexF	11.36211	3.02555	3.755	0.000202	***
sexM	12.25606	3.0223	4.055	6.16E-05	***
age	-0.06913	0.14118	-0.49	0.624697	
addressU	0.14921	0.3967	0.376	0.707049	
famsizeLE3	0.42957	0.33803	1.271	0.204626	
PstatusT	0.15441	0.50217	0.307	0.758649	
Medu	0.11792	0.2242	0.526	0.599248	
Fedu	0.14394	0.19239	0.748	0.454844	
Mjobhealth	0.92644	0.77557	1.195	0.23307	
Mjobother	-0.78181	0.49406	-1.582	0.114441	
Mjobservices	0.46667	0.55345	0.843	0.399684	
Mjobteacher	-0.92238	0.71991	-1.281	0.200942	

Fjobhealth	-0.55394	0.9971	-0.556	0.578865	
Fjobother	-1.13559	0.70857	-1.603	0.109902	
Fjobservices	-0.99389	0.73325	-1.355	0.176131	
Fjobteacher	1.18587	0.89726	1.322	0.187132	
reasonhome	0.16562	0.3842	0.431	0.666666	
reasonother	-0.18025	0.56474	-0.319	0.749782	
reasonreputation	0.44343	0.39874	1.112	0.266853	
guardianmother	0.04995	0.37821	0.132	0.895011	
guardianother	0.86518	0.69022	1.253	0.210855	
traveltime	-0.02432	0.23097	-0.105	0.916209	
studytime	0.60444	0.19893	3.038	0.002554	**
failures	-1.31429	0.23088	-5.692	2.62E-08	***
schoolsupyes	-2.15563	0.4625	-4.661	4.46E-06	***
famsupyes	-0.97932	0.33023	-2.966	0.003225	**
paidyes	-0.10213	0.33113	-0.308	0.757938	
activitiesyes	-0.05332	0.30694	-0.174	0.862181	
nurseryyes	0.02909	0.38009	0.077	0.939041	
higheryes	1.14209	0.74326	1.537	0.12528	
internetyes	0.25513	0.42953	0.594	0.552912	
romanticyes	-0.21106	0.32542	-0.649	0.517029	
famrel	0.02547	0.17	0.15	0.880983	
freetime	0.25506	0.16413	1.554	0.121078	
goout	-0.41367	0.1557	-2.657	0.008243	**
Dalc	-0.06307	0.22951	-0.275	0.783627	
Walc	-0.02551	0.1718	-0.148	0.882052	
health	-0.1676	0.11163	-1.501	0.134152	
absences	0.01222	0.01988	0.615	0.539082	

$$R^2 = 0.9437$$

$$Adjusted\ R^2 = 0.9375$$

$$F: p-value < 2.2e-16$$

根据显著性约简模型，挑选出性别、学习时长、不及格次数、学校补助、外出时长特征进行拟合，结果如下。

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
sexF	11.1551	0.6497	17.169	< 2e-16	***
sexM	12.0266	0.603	19.945	< 2e-16	***

studytime	0.5748	0.19	3.026	0.00264	**
failures	-1.4317	0.2056	-6.963	1.43E-11	***
schoolsupyes	-2.0218	0.45	-4.493	9.26E-06	***
goout	-0.3504	0.1358	-2.58	0.01025	*

$$R^2 = 0.9332$$

$$Adjusted\ R^2 = 0.9322$$

$$F: p-value < 2.2e-16$$

$$MAE = 2.4106$$

$$G_1 = \beta_0 sexF + \beta_1 sexM + \beta_3 studytime + \beta_4 failures + \beta_5 schoolsupyes + \beta_6 goout$$

可以看出，在第一学期，性别为男，学习时间更长，班级过去不及格次数少、学校学习补助低、外出次数少的同学成绩更好，对此有如下解释：

由于数据集的原因，成绩仅录入了数学与葡萄牙语，有较大可能性在中学阶段男性学生对数学更感兴趣，因此成绩会较高，同时在葡萄牙语方面没有较大差距。

学习时间更长，成绩更好，符合直觉。

班级不及格次数，可以代表班级的好坏，所处环境会影响成绩，符合直觉。

学校学习补助低，成绩更好，本文猜测由于获取补助更多的学生家庭条件更差，因此会影响成绩。

外出时间与学习时间的相关系数为-0.06825586，因此外出时间是一个独立的参数，外出时间更长，学习成绩越低，符合直觉。

此外，还对正态性做了检验，结果如下（MAE 值差，R 值好）。

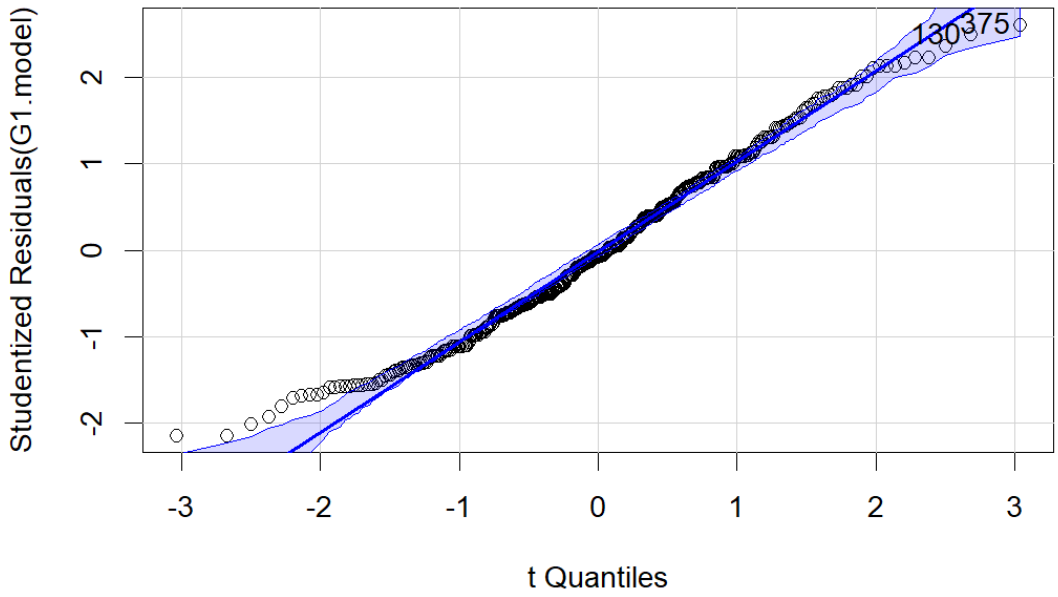
同时对模型进行评估与诊断，发现残差 Q-Q 图分布并不是很好，同时删除了一些异常点，诊断结果如下。

	rstudent <dbl>	unadjusted p-value <dbl>	Bonferroni p <dbl>
199	3.123512	0.0019212	0.75889

同时对线性进行探索（crPlots），发现建议采用因变量方根处理，结果如下。

$$Suggested\ power\ transformation: 0.4696524$$

最终对模型进行探索处理，发现对因变量进行四次方根处理，同时删除 index 为 199 的异常值，对模型进行拟合的效果最好，结果如下。



Coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)	
sexF	1.809044	0.027113	66.722	< 2e-16	***
sexM	1.8498	0.025153	73.541	< 2e-16	***
studytime	0.027064	0.007903	3.424	0.000683	***
failures	-0.06841	0.008877	-7.706	1.12E-13	***
schoolsupyes	-0.08746	0.019368	-4.515	8.43E-06	***
goout	-0.01565	0.005716	-2.739	0.00646	**

$$R^2 = 0.9954$$

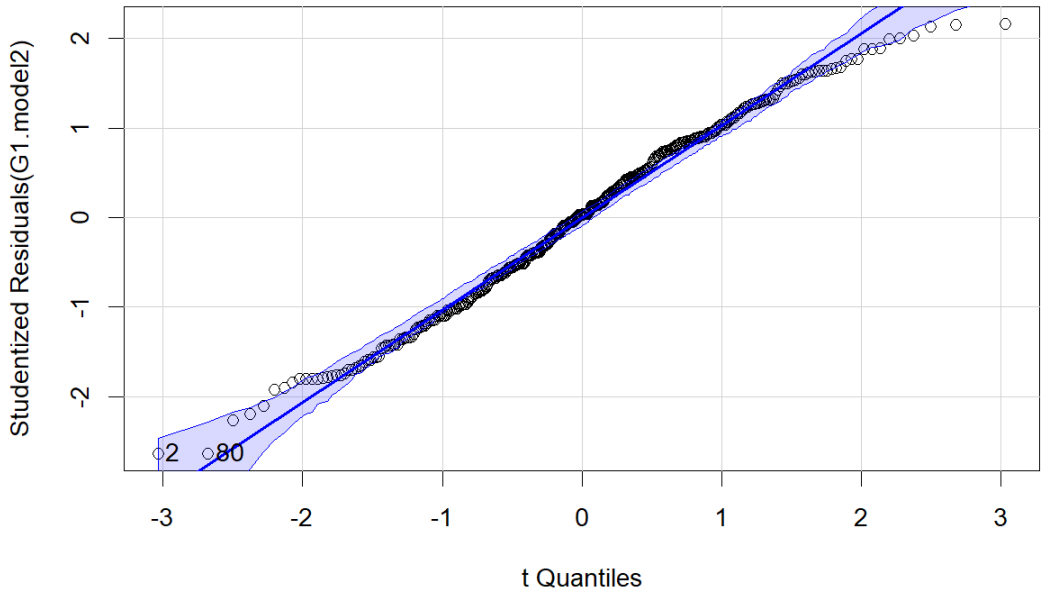
$$Adjusted\ R^2 = 0.9954$$

$$F: p - value < 2.2e - 16$$

$$MAE = 9.1252$$

可以看出，拟合优度很高，但 MAE 值相差很大，同时四次方根较难以解释，因此弃用该模型。

模型残差 Q-Q 图如下。



（三）、 对第二学期成绩的拟合与解释

预测第二学期成绩有两种方法，并入 G1 与不并入 G1，从主观感受上来说，第二学期的成绩受到第一学期成绩一定的影响（相关系数 0.8521181。

与上文流程一致，通过显著性约简变量，得到了如下结果。

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
traveltime	-0.39416	0.14028	-2.81	0.0052	**
romanticno	1.08051	0.41485	2.605	0.00955	**
romanticyes	0.45259	0.4279	1.058	0.29084	
G1	0.95464	0.02949	32.37	< 2e-16	***

$$R^2 = 0.9713$$

$$Adjusted\ R^2 = 0.971$$

$$F: p - value < 2.2e - 16$$

$$MAE = 0.3617$$

Romanticyes 变量是 romantic 二值变量通过独热编码转换而来的变量，可以忽略。

当 traveltime 每增加一个单位时，G1 的平均值就会下降约 0.39 个单位。类似地，当学生 romantic 相信自己没有恋情时，他们的 G1 平均值就会比他们认为自己有恋情的同龄人多约 0.6 个单位，可以得出有恋情会在一定水平上降低学习的成绩。

可以看出，在学习的中期（第二学期），在 G1 影响因素外，是否旅游和是否有恋情都会对成绩有影响。

同时，当选择不并入 G1 时，模型的拟合效果和变量选择会发生较大的差距，选择结果如下。

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
sexF	11.2197	0.7513	14.935	< 2e-16	***
sexM	12.2446	0.6972	17.562	< 2e-16	***
studytime	0.5241	0.2197	2.386	0.01751	*
failures	-1.6473	0.2378	-6.928	1.77E-11	***
schoolsupyes	-1.2095	0.5203	-2.325	0.02061	*
goout	-0.4346	0.1571	-2.767	0.00593	**

$$R^2 = 0.91$$

$$Adjusted\ R^2 = 0.9086$$

$$F: p-value < 2.2e-16$$

$$MAE = 2.4357$$

可见模型在拟合优度和 MAE 上和并入 G1 有较大差距，同时变量也有较大不同，因此在下文的 G3 最终预测上采用并入 G1 与 G2 的方法。

至于多重共线性，由于没有加入截距项，所以 VIF 值有极大概率不准确，因此本文并不考虑多重共线性假设。

（四）、 对第三学期（最终）成绩的拟合与解释

依旧是与前文一致的约简手段并且处理了一些异常值（索引值为 265,342,141,297,260,311,335,344,334,338），最终得出的模型结果如下。

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
sexF	-1.92695	0.39769	-4.845	1.84E-06	***
sexM	-1.97427	0.41214	-4.79	2.39E-06	***
famrel	0.19136	0.07978	2.399	0.0169	*
G1	0.10141	0.04074	2.489	0.0132	*
G2	1.0005	0.03585	27.904	< 2e-16	***

$$R^2 = 0.9856$$

$$Adjusted\ R^2 = 0.9854$$

$$F: p - value < 2.2e - 16$$

$$MAE = 1.3298$$

在学习的后期（即第三学期），除了 G1 与 G2 的影响因素之外，值得注意的是女性学生比男性学生第三次成绩高，说明女性学生在学习后期进行发力。

同时家庭关系质量（famrel）越高，成绩越好。说明在学习后期，更加追求环境的稳定性，保持原有的学习状态，发挥稳定更关键。

三、 模型总结与反思

对于多变量的线性模型，本文采用显著性约简与随机森林回归剪枝（置于附录）的方法进行处理，均得出了较好的拟合优度/方差解释比，能较好地解释影响成绩的主要因素和对成绩进行预测。

然而，这样的方法也存在一些问题：显著性在不同特征组合的情况下显著性并不相同，同时不能仅依靠拟合优度对模型进行选择。考虑实际情况，本文通过拟合优度与 MAE 指标共同选择模型。

通过模型的拟合，本文发现在前期性别为男，学习时间更长，班级过去不及格次数少、学校学习补助低、外出次数少的同学成绩更好。在中期时，在前期努力的基础上，外出次数少和没有恋情的同学成绩更好。在后期时，在前中期努力的基础上，性别为女，家庭关系质量越高的同学成绩更加优异。

附录

1. 随机森林选择

即采用树模型对数据进行回归操作，并且根据特征重要性进行交叉验证辅助选择最优子集进行回归。

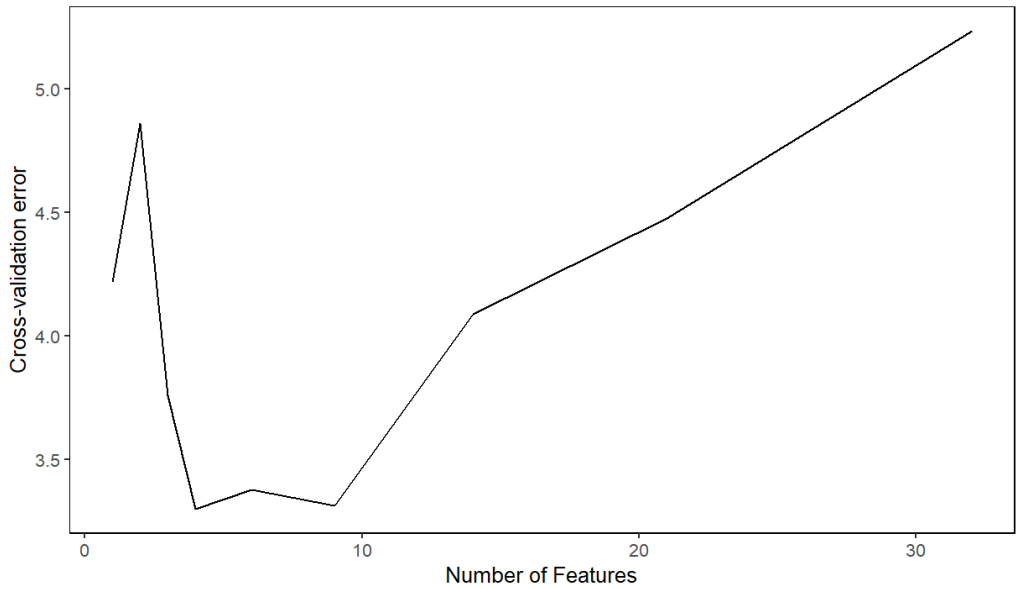
全子集随机森林回归如下。

```
Call:
 randomForest(formula = G3 ~ ., data = train, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 10

          Mean of squared residuals: 3.280447
                    % Var explained: 84.74
```

根据 IncNodePurity 指标对特征进行选择 ， 并进行 10 折交叉验证。结果如下。

	Group.1 <dbl>	x <dbl>
1	1	4.215983
2	2	4.860694
3	3	3.757395
4	4	3.297042
5	6	3.377031
6	9	3.310169
7	14	4.088172
8	21	4.474116
9	32	5.235438



从图中可以看出，n=4 为效果最好的点，选取特征为 G2+G1+absences+failures。结果如下。

```
Call:
 randomForest(formula = G3 ~ G2 + G1 + absences + failures - 1,      data = data_G3_2, importance
= TRUE, ntree = 5000)
               Type of random forest: regression
                     Number of trees: 5000
No. of variables tried at each split: 1

          Mean of squared residuals: 2.556854
                    % Var explained: 86.25
```

方差解释比为 86.25%。