Statistics and Application 统计学与应用, 2016, 5(3), 232-236
Published Online September 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.53023

Hans汉斯

The Analysis on Influence Factors of College Students' Failing in Tests Based on Generalized Linear Model

Yingqin Nie, Jian Zhang, Jie Gao, Yu Zhang

College of Science, China University of Mining & Technology (Beijing), Beijing
Email: 547798905@qq.com

Received: Aug. 18th, 2016; accepted: Sep. 3rd, 2016; published: Sep. 9th, 2016

Abstract

College students' failing in tests represents intensively the students' learning situation in the college. This paper identified a number of factors that have a significant impact on the total number of college students' failing in tests by statistical modeling and gave some advices that improve the college students' academic performance. The phenomenon over discrete came out in the process of Poisson regression, so we used a negative binomial regression model.

Keywords

College Students' Failing in Tests, Poisson Regression Model, Negative Binomial Regression Model

基于广义线性模型的大学生挂科影响因素分析

聂映芹，张　建，高　洁，张　雨

中国矿业大学(北京)理学院，北京
Email: 547798905@qq.com

收稿日期：2016年8月18日；录用日期：2016年9月3日；发布日期：2016年9月9日

摘　要

大学生挂科情况是学生在校学习情况的集中体现，本文通过统计建模确定了一些对大学生挂科门数具有

文章引用: 聂映芹, 张建, 高洁, 张雨. 基于广义线性模型的大学生挂科影响因素分析[J]. 统计学与应用, 2016, 5(3): 232-236. http://dx.doi.org/10.12677/sa.2016.53023

显著影响的因素，给出了一些提高大学生在校学习情况的意见。统计建模中鉴于泊松回归模型存在过离散现象，采用了负二项回归模型。

关键词

大学生挂科，泊松回归模型，负二项回归模型

1. 引言

大学生在校学习情况是学生素质的集中体现，关系着学生的就业、求职等方面。挂科门数是在校学习情况的一种数值度量。常见对于大学生挂科影响因素的定量分析都是基于 logistic 回归模型进行的，其中大学生挂科与否为因变量。以大学生挂科门数作为因变量，基于计数模型进行定量分析的并不常见。

本文所要研究的学生挂科门数为计数变量。计数变量是指取值为非负整数的变量，如保单的索赔次数、家庭的孩子数、旅游景点的访问人数等。针对计数变量的较为常见的统计建模方法为对数线性模型(泊松回归模型)，但过离散现象出现时该方法已不适宜，此时负二项回归模型比较适宜。

通过统计建模的定量分析可以数值化的确定学生挂科情况的影响因素，进而可以有针对性的通过一些措施的实施降低大学生挂科门数，提高大学生在校学习成绩。

2. 模型的介绍

无论针对二元变量的 logistic 回归，还是针对计数变量的泊松回归、负二项回归都是广义线性模型的特例。

广义线性模型的模型假设由三部分构成[1] [2]：

1) 因变量 Y_i，$i=1,2,\cdots,n$ 相互独立且服从离散指数族分布，其密度函数形式为

$$f\left(y_i,\theta_i,\phi\right)=\exp\left(\frac{\theta_i y_i-b\left(\theta_i\right)}{a\left(\phi\right)}+c\left(y_i,\phi\right)\right)$$

其中 θ_i 为自然参数，与 Y 的均值有关；ϕ 为离散参数，与 Y 的方差相关，ϕ 已知时，该分布族又称指数分布；$a\left(\phi\right),b\left(\theta_i\right),c\left(y_i,\phi\right)$ 为已知函数。

2) 存在基于自变量 $X_i=\left(X_{i1},X_{i2},\cdots,X_{ip}\right)$ 的线性预测 $\eta_i=X_i\beta^{\mathrm{T}}$ 与因变量 Y_i 的期望相关，其中 $\beta=\left(\beta_1,\beta_2,\cdots,\beta_p\right)$ 为未知参数。

3) 存在单调可微联系函数 g 满足 $g\left(u_i\right)=\eta_i=X_i\beta^{\mathrm{T}}$，其中 $u_i=E\left(Y_i\right)$，当 $g\left(u_i\right)=\eta_i=\theta_i$ 时，相应的联系函数 g 称为自然联系函数。

广义线性模型的建模过程主要包括基于迭代重加权数值算法的参数估计、基于似然比检验的拟合优度、基于渐近分布的假设检验、基于 Deviance 残差的残差分析等，McCullagh [3]给出了相应的理论说明和应用实践；Venable [4]给出了基于 GLM 软件包的相应 R 软件实现过程。

泊松回归模型的模型假设一般形式如下：

$$\begin{cases}f\left(y_i,\mu_i\right)=\exp\left(\dfrac{y_i\log\mu_i-\mu_i}{1}-\log y_i!\right),i=1,2,\cdots,n\\ g\left(\mu_i\right)=\ln\left(\mu_i\right)=\eta_i=X_i\beta^{\mathrm{T}}\end{cases}$$

以上模型中假定因变量 y 服从泊松分布，要求离散参数 $\phi=1$，即因变量的期望与方差相同，即 $E\left(y\right)=Var\left(y\right)$。该假设在实际数据拟合时常常得不到满足，较常见的情形是：$Var\left(y\right)>E\left(y\right)$，即离散

参数$\phi>1$，该情形称为过离散现象，此时因变量已不服从假定的泊松分布，基于以上模型的建模过程无效。面对这种情形，较为常见的统计建模过程是拟合负二项回归模型。负二项回归模型的一般形式如下：

$$\begin{cases} f\left(y_i,\mu_i,\theta\right)=\dfrac{\Gamma\left(y_i+\theta\right)}{\Gamma\left(\theta\right)y_i!}r_i^{y_i}\left(1-r_i\right)^{\theta},r_i=\dfrac{\mu_i}{\mu_i+\theta},i=1,2,\cdots \\ g\left(\mu_i\right)=\ln\left(\mu_i\right)=\eta_i=X_i\beta^{\mathrm{T}} \end{cases}$$

3. 大学生挂科影响因素建模分析

本文研究数据取自某重点工科高校 2012 级不同专业不同班级大学生前五个学期课程考试成绩和学生的一些基本资料。由于部分学生转专业、退学、休学等情况的出现造成了数据的不完整，本文将相应数据进行了删除；对采用等级级打分形式、学生全部及格的相应课程，本文也进行了相应数据的删除。最终在数据收集整理过程中得到的样本数据的样本容量是 557；其中男女比例为 116:441；党团员比例为 8.6%；学生的平均年龄为 18.9 岁。通过变量之间的相关性分析和一些基本常识，本文初始选定年龄T_1、性别T_2、政治面貌T_3、高数成绩T_4、概率成绩T_5、C 语言成绩T_6作为自变量，因变量为相应大学生五个学期课程挂科门数Y。

由于挂科门数Y为计数数据，故本文在进行相应分析时，初始模型为泊松回归模型。在变量选择过程中，无论是向前(向后)还是逐步变量选择的过程得到的结论是一致的：在显著水平$\alpha=0.05$下，年龄T_1与政治面貌T_3对挂科门数无显著影响，剩余变量影响显著。故本文选定性别T_2、高数成绩T_4、概率成绩T_5、C 语言成绩T_6作为最终自变量。

如上所述，本文采用性别X_1、高数成绩X_2、概率成绩X_3、C 语言成绩X_4作为自变量，对应的泊松回归模型是

$$\begin{cases} f\left(y_i,\mu_i\right)=\dfrac{\mu_i^{y_i}}{y_i!}\mathrm{e}^{-\mu_i},i=1,2,\cdots,n \\ \ln\left(\mu_i\right)=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i3}+\beta_4X_{i4} \end{cases}$$

基于 R 软件 GLM 软件包以上统计模型的拟合结果如表 1 所示。

从表 1 可以看到，该模型的残差离差 681.9 远大于自由度 556，可见拟合效果较差，过离散现象存在，需要在建模中考虑并解决过离散现象。

计数数据统计建模中的过离散产生根源是多方面的，如模型中尚有没能包含的重要解释变量、或个体事件的发生存在相关性或聚类性、或存在异常影响点、或模型本身指定有误、或数据中存在过多的零计数等等。由于对于泊松回归模型而言，过离散现象的外在表现就是因变量的方差大于期望，而满足方差大于期望且可用于计数数据统计建模的另一常见统计分布为负二项分布，故而负二项回归模型成为过离散现象存在情形下计数数据统计建模的标准方法之一。

基于 R 软件 MASS 软件包中的命令 glm.nb，以上分析过程的负二项回归模型拟合结果如表 2。

由表 2 可知泊松回归模型的 AIC 为 1266，而负二项回归模型的 AIC 为 1206.4，可见负二项回归模型比泊松回归模型确实能更好的拟合观测数据。

进一步由表 2 可知未知参数β的估计为

$$\hat{\beta}=\left(7.9179,-0.7091,-0.0440,-0.0396,-0.0343\right)$$

即性别X_1的回归系数为−0.7091，意味着在保持其它变量不变情形下，性别变量取值为 1 时的挂科门数是性别变量为 0 时的$\exp\left(\beta_1\right)=49.21\%$，意味着平均意义下女生挂科门数约为男生挂科门数的一半；

Table 1. Model fitting based on Poisson regression
表 1. 基于泊松回归的模型拟合

变量名	估计值	标准差	P 值
截距项	7.5449	0.3127	<2e−16
性别	−0.5830	0.14735	7.61e−05
高数成绩	−0.0439	0.0044	<2e−12
概率成绩	−0.0350	0.0039	<2e−16
C 语言成绩	−0.0332	0.0049	9.59e−12
离差残差：681.89	自由度：552	AIC：1266.4	

Table 2. Model fitting based on binomial regression
表 2. 基于负二项回归的模型拟合

变量名	估计值	标准差	P 值
截距项	7.9179	0.4547	<2e−16
性别	−0.7091	0.1955	0.00029
高数成绩	−0.0440	0.0063	2.96e−12
概率成绩	−0.0396	0.0055	8.61e−13
C 语言成绩	−0.0343	0.0068	3.73e−07
离差残差：458.22	自由度：552	AIC：1206.4	

高数成绩 X_2 的回归系数为−0.0440，且具有高度显著性，意味着其它变量保持不变情形下，高数成绩每提高 1 分，学生挂科门数是未提高前的 $\exp(\beta_2)=95.68\%$；高数成绩每提高 5 分，学生挂科门数是未提高前的 $\exp(5\beta_2)=80.2\%$；每提高 10 分，学生挂科门数是未提高前的 $\exp(10\beta_2)=64.33\%$；概率成绩 X_3 的回归系数为−0.0396，意味着其它变量保持不变情形下，概率成绩每提高 1 分，学生挂科门数是未提高前的 $\exp(\beta_3)=96.12\%$；概率成绩每提高 5 分，学生挂科门数是未提高前的 $\exp(5\beta_3)=82.06\%$；概率成绩每提高 10 分，学生挂科门数是未提高前的 $\exp(10\beta_3)=67.33\%$；C 语言成绩 X_4 的回归系数为−0.0343，意味着其它变量保持不变情形下，C 语言成绩每提高 1 分，学生挂科门数是未提高前的 $\exp(\beta_4)=96.62\%$；C 语言成绩每提高 5 分，学生挂科门数是未提高前的 $\exp(5\beta_4)=84.23\%$；C 语言成绩每提高 10 分，学生挂科门数是未提高前的 $\exp(10\beta_4)=70.94\%$。

4. 总结

针对计数数据的常用泊松回归模型拟合过程中过离散现象的时常发生，McCullagh [3]指出：除非有充分理由相信因变量服从泊松分布的假设，否则更明智的做法是建模初始就考虑过离散现象的存在，此时负二项回归模型是一个较为常见的备选模型，但是两种模型都是广义线性模型的特例。

大学生在校期间的挂科情况是学生在校学习情况的集中体现之一，对其影响因素的数值分析是十分有意义的。本文的模型拟合结果表明：

1) 平均意义下女生的挂科门数远小于男生，即女生的平均成绩远优于男生，所以在学习上可进行男女搭配模式的帮扶小组，授课过程中应多关注一些男生同学的听课注意力，多对男生进行提问，这对提高大学生在校学习情况将会有一定的作用。

2) 除性别这个定性变量之外，其他 3 个自变量均为课程考试成绩：高等数学，概率论与数理统计，

C 语言。拟合结果表明如果这 3 门课程的考试成绩得以提升，那么挂科门数就会有相应大幅比例的下降，且高等数学效果最为显著。高等数学的效果显著一方面是由高等数学分上、下两学期且知识点较多所决定的，另一方面是由高等数学是工科后续学习课程的基础的地位所决定的。一定要充分重视以高等数学为代表的这三门课程的学习。对授课教师来讲，在保证授课课时的前提下，增加一些辅助学习环节，例如课下的答疑课、章节习题课、知识结构讲解课等；采取一些方法提高学生的学习兴趣，例如讲解一些与课程相关的数学史的知识，讲解一些与学生所学专业相关的课程知识等；转变一些教学方法，例如在期中考试过程中增加口试的过程，深入了解学生的课程掌握情况，学生对课程学习的疑惑等；例如平时作业不再从课后习题指定而是任课教师自己出题等。对学生来讲，在上课认真听讲下课自行完成作业的同时，应尽可能抽时间看一些指定教材之外的教材，相同知识点不同的讲解方式易于接受；应在学习章节知识点的同时自主去回顾构建相应课程的知识结构体系。

基金项目

本文受北京市科研创新训练项目“广义线性模型的理论与应用”(项目编号：C201507002)的支持。

参考文献 (References)

[1] Lindsey, J.K. (1997) Applying Generalized Linear Models. Springer, New York, 18-20.

[2] 陈希孺. 广义线性模型(一) [J]. 数理统计与管理, 2002, 21(5): 54-61.

[3] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2nd Edition, Chapman & Hall, London, 21-44. http://dx.doi.org/10.1007/978-1-4899-3242-6

[4] Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics with S. 4th Edition, Springer, Berlin, 183-208. http://dx.doi.org/10.1007/978-0-387-21706-2