第 5 卷◆第 3 期◆版本 1.0◆2022 年
文章类型：论文|刊号（ISSN）：2630-4686 /（中图刊号）：380GL020

大数据时代经管类统计学教学改革探究

孙婧
上海商学院 商务经济学院
DOI:10.12238/er.v5i3.4593

[摘　要] 本文总结了经济管理类专业统计学教学中存在的一些问题,提出结合R语言进行实验教学和引入综合性、开放性实践教学项目的改革思路和方法。在教学中,要重视发挥学生的创造力,着力培养学生运用所学知识的实践能力。
[关键词] 大数据；统计教学；R语言；实践教学
中图分类号：G424.1　文献标识码：A

Exploration on the statistics teaching reform for the Economics and Management in the Era of Big Data

Jing Sun
School of business and economics, Shanghai Business University

[Abstract] In this paper, the author summarizes some problems existed in Statistics Teaching in the Economics and Management. On top of that, the author presents some reform ideas and methods , such as combining R language with experiment teaching, bringing comprehensive and open practical teaching project. In the teaching process, we should pay attention to the development of students' creativity and focus on cultivating their' practical ability to apply the knowledge what they have learned.
[Key words] Big Data; Statistics Teaching; R Language; Practice Teaching

引言

“大数据”的概念最早是由硅图公司(SGI)的首席统计学家John R. Mashey 于1998年提出的。McKinsey Global Institute的定义是：一种规模很大,在取得、储存、管理、分析方面远超出传统数据库软件处理能力范围的数据集合。其具有数据规模(Volume)海量、数据流转(Velocity)快速、数据类型(Variety)多样和数据价值(Value)密度低的特征。

统计学是一门关于数据的获取、整理、展示、分析、解读的方法论学科,是经济、管理、财务、金融等专业必修的一门专业课程。大数据时代的到来,使传统的统计学理论和分析工具、分析方法面临挑战,继而就会影响到统计学的教学,统计学要与时俱进,必须要进行教学内容的重构和教学方法的创新。

朱建平[1]认为大数据时代,人们可以更加主动地利用数据,这就需要将传统的统计技术与现代计算机技术紧密结合。孟生旺、袁卫[2]认为,在大数据时代背景下,统计学应该从“树立大数据思维、加强计算机统计计算和编程能力培养、注重实践训练、结合特定专业”等方面进行改革。周茂袁[3]认为在大数据时代,统计学应该增加R语言、数据挖掘等内容。

1 经管类统计学教学现状分析

1.1教学内容不合理

目前,在经济管理类的统计学教学中,还没有将统计思想的理解、统计方法的掌握、实际问题的解决作为教学重点,仍旧用讲数学的思维逻辑讲解统计学,重点放在统计学的一些基本概念和理论内容上,把统计学教学看成是统计概念的讲述和公式的推理,忽视了对统计方法的介绍,对统计软件讲解和实践往往也是一带而过,这与统计学的教学目的相违背。

大数据分析需要加强统计学与计算机的合作,解决经济领域的实际问题。一些证明和统计计算,点击计算机就能解决,在教学中应该弱化。应该教给学生的是：用这种方法的理由是什么？该方法应用场景是什么？使用这种方法时,有没有前提条件？怎么判断前提条件是否满足？满足到何种程度？…….

大数据时代对统计学的理论与分析工具产生了诸多的影响和冲击,改善教材、重构教学内容、改革教学方法具有必要性和紧迫性。

1.2实践教学缺乏有效性、综合性、开放性、先进性

目前,高校对实践教学环节越来越重视,对人才培养强调的是“应用型”。在大数据时代,统计学教学中的薄弱环节也是实践教学。主要体现在三个方面：(1)实践教学缺乏先进性。在现有的实践教学中,提到的和使用的大多还是传统的统计分析软件,如Excel、SPSS和Eviews等,使得学生无法围绕统计学习的主线,系统化地学好一个统计软件,而是碎片化式地应付各种各样的软件。大数据时代,数据存在海量、多样、高速的特征,需要多样化的统计模型和方法,需要学生具有统计计算和编程能力,用这些软件已经无法满足教学的要求。(2)实践教学缺乏有效性。现有的实践教学内容设计以验证统计原理为主,多是利用一些适应模型的特殊数据来“证明”模型的正确性和有效性,出发点并不是数据,这种思维方式与大数据时代的要求南辕北辙,统计模型应该是用来拟合真实数据的。(3)实践教学缺乏综合性、开放性。实践教学内容脱离计算机技术发展和企业的需求,缺少开放性、设计性和综合性实验,不利于发挥学生的主动性的和培养他们的创造力。

1.3学生的学习兴趣偏低

目前,在设置经济管理类专业培养计划时,是把微积分和概率论知识作为基础放在统计学学习之前,统计学放在大二或大三开设,其课程地位处于数学理论付诸实践应用的环节上。如果学生在微积分、概率论等课程的学习中已经出现困难,则很可能在统计学课程中形成一种负面累积效应,在一些学生心目中,统计学还是概念多、公式复杂的代名词。看着教材上繁复的数学公式,如果不能透过现象看本质,就容易把统计学与数学划等号,从而产生对课程的抵触情绪、畏惧心理,影响其学习兴趣。

在经济管理类专业的课程设置当中,后续与统计相关的数据分析和数据挖掘的课程较少,如果统计学教学延续数学教学的思维模式,不以传递大数据统计思维为出发点,缺乏有效的实践教学,学生就很难具备今后经济管理岗位要求的数据分析处理能力,学习兴趣也会较低。

2 大数据时代统计学教学改革思路

2.1重构教学内容,传递大数据统计思维

在传统的统计学教学中,只要求学生能够利用统计学理念来解决“小数据”问题。大数据时代是以数据为中心的时代,需要让学生改变其对样本的认识,对不确定性的认识;需要让学生学习新的数据处理与分类的方法,理解结构化数据与非结构化数据的对接；需要让学生建立相关分析与因果分析并重的思想,使他们具备构建数据仓库、数据预处理、数据挖掘等大数据能力。

2.2结合R语言,加强实验教学

R语言属于一个免费、自由、源代码开放的软件,统计计算、可视化和统计建模功能强大,也是最新统计方法发布的主要平台。与多数统计软件相比,R语言包含很多最新统计方法的实现方案,而且它的编程较为简单,只要有统计理论支撑,不需要高深的计算机基础知识,也很容易入门,非常有利于培养学生的编程能力和知识更新能力。近几年,伴随着数据仓库、数据分析与挖掘、商业智能以及开源云计算架构的兴起,R语言软件在社会上的影响力迅速提升,在高校的教学中也逐渐被重视,特别是在统计学教学中。

XX网站关于数据分析岗位的职责及要求

职责	要求
1、参与团队合作,完成产品课题, 2、运用专业技术手段解决课题(通过数据分析) 3、撰写数据分析报告呈现给决策者并解读 4、收集新数据 5、对数据收集方法进行优化 ……	1、会使用数据分析、挖掘工具软件(例如：SAS、R、Python) 2、熟练掌握数据库技术和 SQL 3、熟悉神经网络、逻辑回归、决策树、聚类等建模方法并能使用其中一种或几种建模 ……

统计分析的过程大致可以分为五个阶段：收集数据、处理数据、描述统计、建模、模型的评估与应用。而统计学的教学工作也应该按照这五个阶段的顺序依次展开,在上述五个阶段中,R语言主要应用在处理数据、描述统计、建模这三个阶段,处理数据阶段包括数据的预处理、数据的清洗整合等方面。描述统计阶段包括对数据的分析和可视化等。建模阶段主要包括一些统计方法的应用,如回归分析、方差分析、主成分与聚类分析、时间序列分析等。如果强调理论与应用并重,强调训练与大数据分析相关的计算机技能,R语言的优势非常突出。而且只使用一种软件与课堂教学融会贯通,更能提升学生的学习兴趣以及对R语言的掌握运用。

在统计教学过程中,R语言的优势还在于它能进行可视化教学。它能够将抽象的统计学概念转化为直观的图形或者函数,把各种统计数据通过图形直观的展示给学生,使抽象的理论变得形象化,使统计学学习变得生动易懂。

基于R语言的统计教学案例[4]

例：一元线性回归模型

R自带的数据包MASS库中有一个名为“Boston”的数据集。以该数据集为例来说明一元线性回归的统计教学。这个数据集展现是Boston郊区的住宅价格,它有14个变量和506行观察值。假定我们选择其中一个变量rm(每套住宅的平均房间数)作为自变量,另一个变量medv(住宅价格的中位数)作为因变量,用一元线性回归方法来预测住宅价格。

```
library(MASS)
data(Boston)
names(Boston)
[1]"crim" "zn" "indus" "chas" "nox" "rm" "age" "dis" "rad"
```

[10]"tax" "ptratio" "black" "lstat" "medv"

lm()是R中执行线性回归算法的工具。将计算结果存放在lm1这个对象中。编程如下：

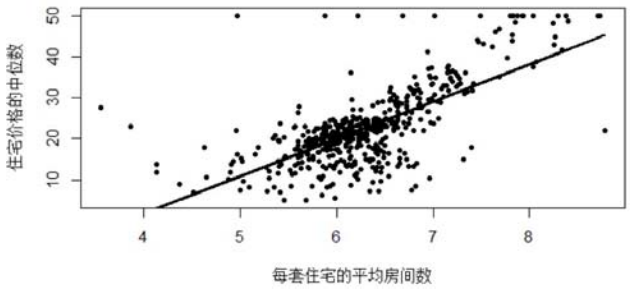
```
lm1<-lm(medv~rm,data=Boston)
```

可以使用summary()函数来展示得到的具体模型：(1)回归系数(斜率和截距)。(2)回归系数的标准误差、判定系数、P值。用以评估模型的显著性。

```
summary(lm1)
Call:
lm(formula=medv~rm,data=Boston)
Residuals:
Min  1Q  Median  3Q  Max
-23.346  -2.547  0.090  2.986  39.433
Coefficients:
Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -34.671  2.650  -13.08  <2e-16 ***
rm  9.102  0.419  21.72  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error:  6.616  on  504 degrees of freedom
Multiple R-squared:  0.4835, Adjusted  R-squared:  0.4825
F-statistic: 471.8 on 1  and 504 DF,  p-value: < 2.2e-16
```

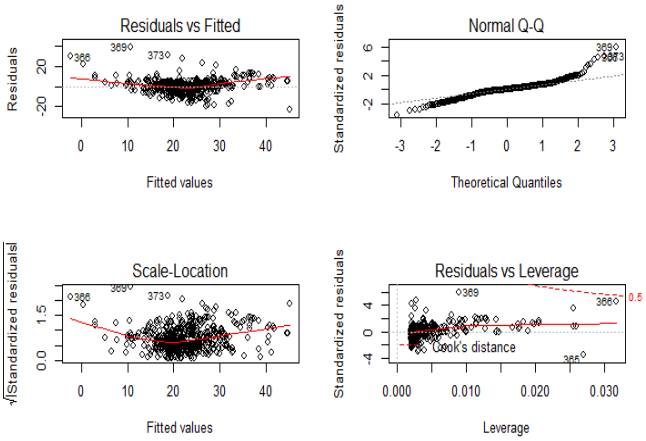
使用R的绘图功能，还可以将该模型可视化。

```
attach(Boston)
plot(rm,medv,pch=20,xlab="每套住宅的平均房间数",ylab="住宅价格的中位数")
lines(rm,lm1$fitted,lwd=3)
```



我们还可以对模型进行诊断。R的plot( )函数提供了一套四个图表，可以分别展示残差—拟合值(Residuals vs Fitted)、标准化残差Q-Q图(NormalQ-Q)、比例尺定位图(Scale-Location)和残差—杠杆比例(Residuals vs Leverage)的散点图。可以根据这些图表来对模型进一步分析。例如：我们可以看看残差是否有明显的模式、是否符合正态分布。一些异常值的点在图中标注了出来(如：366、369、373)，这些点是后续需要进一步检查的点。

```
par(mfrow=c(2,2))
plot(lm1)
```



从一元线性回归模型的教学案例可以看出，把R 语言引入统计学课程教学环节，弱化模型的数学推导过程，(对于多元统计分析，数学推导过程甚至是在课堂上难于完成的)把注意力集中在模型的思想和应用层面，引导学生利用统计软件、统计方法来处理数据，同时用可视化的图表解释统计问题大大增强了学生对抽象的数学理论的理解。以R语言来辅助教学，强调统计的应用层面，如模型设定、模型输出、模型检验和评估、模型结果的解释也为学生将来能更快更好地融入数据分析、数据处理等岗位的工作打下了坚实的基础。

2.3以应用为本，改革实践教学项目

实践教学缺乏创新性、实用性、有效性是统计学教学面临的另一问题，应该引入一些综合性实践教学项目。综合性实践性教学项目的选题可以改编来自社会、经济和管理科学等方面的实际问题，有充分的灵活性、开放性供学生发挥创造能力。在完成实践性项目的过程中，着重引导学生学习以统计分析软件为计算工具，以统计分析方法对批量数据建模和检验评估的统计思维。

实践性教学项目

例：哪些人最有可能欠钱不还？(基于网贷平台的贷款数据分析)

第一步获取数据。在这一阶段，教学重在引导学生学习多渠道收集数据、如何围绕研究主题收集数据的方法。例如：读取CSV、Excel文件、从SQL数据库中读取数据、从网站中抓取数据、从网络中下载数据集等。

第二步理解数据、处理数据。收集到的数据往往是大量的、多维度的，也存在数据不一致、数据缺失等情况。在这一阶段，

第 5 卷◆第 3 期◆版本 1.0◆2022 年
文章类型：论文|刊号（ISSN）：2630-4686 /（中图刊号）：380GL020

教学重在引导学生学习原始数据的清洗、将原始数据转换成合适的数据格式的方法。例如：数据采样、修正变量名、创建新变量、重塑数据集、处理缺失数据、特征缩放、降维等。

第三步探索性数据分析、筛选最优变量。在这一阶段，教学重在引导学生学习数据统计、分布、密度、检查所有的因子变量、探索性可视化、解释性可视化等，充分认识数据，更进一步对数据进行提纯或转换，启发建模思路。

例如：1. 信用卡使用情况与贷款状态的关系？

2. 借款人是否有房屋与贷款状态的关系？

3. 客户的职业、月收入、年收入与贷款状态的关系？

4. 客户5年内违约次数与贷款状态的关系？

单变量跟贷款状态的关系、多变量跟贷款状态的关系等。

第四步建模，做预测分析。在这一阶段，教学重在引导学生学习回归与分类的算法。建立数据的训练集与测试集、挑选模型、评估模型的预测水平的方法。

在实践性教学的过程中，教师只根据项目的复杂程度进行启发式讲解，弱化教师的主导，充分调动学生学习的积极性和主动性。将学生进行分组，自主选择题目，课下进行数据收集、数据处理、方案选择等环节，课上进行报告和讨论，在解决问题的全过程中加强学生对统计方法的理解和运用，以“项目引领—收集数据—探索数据—算法建模—优化决策”的思想为主线，积极创设问题情境，通过问题情境的设计，使学生由“被动学统计知识”变成“主动训练统计思维”，极大地提升学生处理实际问题的综合能力，激发学生的求知欲和创造力，达到学以致用。

3 结语

统计学的生命力在于应用。在大数据时代，收集海量数据的难度降低了，但各行各业对数据的重视程度提高了，对数据的挖掘整理能力和敏感性分析的要求越来越高。在大数据的背景下，统计学的思维模式、定义、作用都不同于传统统计。因此，将大数据统计的思想和方法及时纳入统计学教学是非常必要的。同时，统计学也应该与计算机紧密结合，数学为统计学提供了坚实的理论基础，计算机则可以使统计分析更加简单快捷，并能够解决庞大复杂的数据处理问题。以大数据时代为契机，改革统计学教学方法，重构经管类统计学教学内容，培养具有现代统计技术、计算机数据挖掘技术和经管专业知识的复合人才。

[参考文献]

[1]朱建平,张悦涵.大数据时代对传统统计学变革的思考[J]统计研究,2016,33(2):3-9.

[2]孟生旺,袁卫.大数据时代的统计教育[J].统计研究,2015,32(4):3-7.

[3]周茂袁.大数据时代统计学专业教学改革的初步探索[J].教育教学论坛,2015,(9):105-106.

[4](美)古铁雷斯(DanialD.Gutierrez)机器学习与数据科学基于R的统计学习方法[M].北京:人民邮电出版社,2017.

作者简介：

孙婧(1975--),女,汉族,上海人,硕士,上海商学院,副教授,研究方向：应用统计、数据可视化与商务智能。