**基于线性回归对考研人数的分析和预测**

**1 绪论**

**1.1研究背景**

从教育部公布的数据得知,2019年考研报名人数290万人，较之上一年度的238万增长了52万人。2020年考研报名人数341万人，比上一年度的290万多了51万。2021年考研报考人数377万人，比上一年度增长了36万人。

随着社会科技的高速发展，考研大军越来越强大，从教育部获悉，2020年的考研对比往年延迟一周，改到2020年12月27到2020年12月29日（部分专业考试超过3小时的在23号考），在全国研究生人数调查报告中，从2014年开始中国研究生报名人数不断上升，而且涨幅很大，21年首次突破400万人。对于这几年掀起的“考研热”，今年的考试题目已经在与高考的考试格式靠齐，不少人认为考研已经成为了“第二次高考”。对于这种现象，其背后隐藏着很多因素，在最近的一个调查报告中显示，调查中32.6%的大学毕业生会选择直接就业，41.2%的学生会选择读研或者读博，27.2%的学生会选择出国留学。

这个现象到底是好还是不好？一般有以下两个观点，部分人认为好处是，人们都愿意去学习更深层次的知识，培养自己，以后成为国家的顶尖科技技术人才，但是同时，也有部分人是担心自己的本科生学历“不够”，因为社会上普遍存在的“唯学校”、“唯学历”这种用人导向，导致大家努力往名校挤，盲目考研，跟风大家，不知道读研的目的是什么。

**1.2研究目的与意义**

（1）考研人数逐年不断增多的大致原因

①教育资源与社会状态:社会的发展迅速必然导致对高学历人才的需求，各个城市对人才引进的策略中也可以看出大学生和硕士研究生的差异。

九年义务教育的普及，人们普遍受教育程度上升，本科扩招的情况下，大学生的资质已经满足不了社会需求，而且中国人数不断上升，退休的岁数更加晚，找工作更难，导致人们对于研究生学历的需求多了起来。

②初始工资和就业能力：大部分的公司对于研究生和本科生的初始工资和要求不同，而且在升职过程中，有些要求了不是研究生不能升职，还有一些公司的某些部门直接表明不是研究生不收，或者‘研究生优先录取’，如果没有研究生学历，将无法通过公司的第一关。

③考博的需求和科研能力：对于高学历人才的培养，提高国家的整体教育水平，或者是培养高端人才，都十分的有意义。国家对于研究生的扩招计划，在一定程度上反应了国家对于研究生和硕士的需求，越来越多的人可能会选择钻研科研和学术，这些人以后都可能会是国家的顶尖的人才。

（2）研究意义

上面分析了考研人数增多的原因，下面讨论一下影响考研人数增长的因素。将从四个大方向上对于考研人数增长原因进行分析。首先是社会飞速发展，人们的生活质量不断提升，所以可以用GDP量化这个数据，从国内生产总值能在一定程度上反映社会的发展；其次通过录取人数的增多，能够量化国家对于研究生人才培养计划的提高，近几年国家对于研究生的补助越来越多，而且国家缺乏高科技人才，所以这几年大力鼓励高素质人才的培养；然后就是可以利用失业率在一定程度上反映社会压力，社会上对于所处院校的高要求，对于学历的高要求，无疑对于本科生会产生不小的压力；最后一个层面是利用在校大学生的人数对于学生之间的竞争压力进行量化，不断增长的毕业生人数，在2018年突破800万人数，在走出社会的最后一步，就导致了不少人会选择继续考研。综上所述，通过对于生活质量的提高，国家对于考研政策的优化，社会给本科生给的压力，本科生与本科生的竞争压力，这四个大方面对于考研人数的影响，通过分析来得到最后的结果，然后提出几点建议。

对于笔者个人的意义，在考研期间遇到的一些问题，得到了学长学姐还有老师的建议，成功上岸后，希望能够通过这个项目能对于以后的学弟学妹给出些许建议，希望学弟学妹们能够更加了解考研的目的，不要盲目的追求“考研热”，同时这个课程是研究生期间需要学习的入门课程，通过一段时间的学习，能够让笔者对于机器学习的内容有入门的水平，在研究生期间能够更加的得心应手。

**1.3国内研究现状**

对于考研人数的原因分析，目前国内在此方面的研究较少，早一点就是对于2000-2010年的数据分析，并且存在的设计都是用现有的软件，matlab，spss，Eviews，直接将数据输入，就得到了回归模型。

面对越来越多的数据，计算机行业人员需要做的是，通过数据发现数据中隐含的信息，然后将数据通过一些模型和算法的处理，达到预测的效果，数据

分析师现在无论是在国内，还是国外都是比较火的行业之一。

**1.4本文研究内容与内容框架**

本文通过多元线性回归模型，以报考录取人数，在校大学生毕业生人数，GDP，失业率四个值作为自变量，考研报考人数为因变量，对于每个元素所占比重大小进行分析，并以2000-2019年二十年的数据按照7:3的训练集和测试集，对于模型评分，并按照前十年的数据对于2020年的考研报考人数进行预测，与真实值作比较，看出入多少（2020年开始有疫情，影响较大）。

本文的余下部分安排如下：

第二章介绍了本文多元线性回归模型的模型建立，以及梯度下降的相关知识基础，例如多元线性回归模型的产生和多元线性回归模型的公式，梯度下降的理论解释和梯度下降的公式，以及公式中学习率和迭代次数的选取。

第三章重点介绍本文的设计过程，设计过程分为三部分，首先通过自变量和因变量的关系判断用哪种模型，然后通过模型的分析对损失值做最小化，最后用数据对未来的数据进行预测，并对模型评分。

第四章对设计的算法进行实验论证，检验算法的可行性与正确性。

第五章总结，将本算法的优缺点列举出来，然后对于未来研究方向做出一点规划。

**2 线性回归算法相关知识基础**

**2.1线性回归模型**

**2.1.1 一元线性回归模型**

一元线性回归模型是只有一个因变量和一个自变量的线性相关关系的方法，一个数值往往受到很多原因的影响，但是其中的一个起到决定性作用，就可以用一元线性回归模型分析。公式是：

（1）

与初中所学的直角坐标系中的直线方程式类似,x是自变量，y是被解释变量，a是回归系数，b是截距。

2.1.2多元线性回归模型

现实生活中，一个数值的变化，往往与很多因素相关，例如房价的数值不仅与房子所处的位置（离市区的距离），还与房子面积，房子几室几厅等相关因

素有关，这时候就需要用到多元线性回归模型对以上这些原因分析，来判断哪个元素占比更大。

多元线性回归模型的一般形式为:

Y= （2）

其中、是k+1个未知参数，是回归常数，称为回归系数，y是被解释变量，、到是k个可以控制的一般变量，称为解释变量当n为1的时候就是一元线性回归模型，当n>=2时，就是多元线性回归模型，一般与真实情况会有误差，需要将误差最小化，就是让误差对每个β求偏导，就能得出拟合程度较高的回归模型了。

**2.1.3非线性回归**

线性回归，自变量和因变量之间存在着线性关系，非线性回归，就是自变量和因变量之间不存在线性相关关系，也称线性无关。处理方法一般通过数学公式，将非线性回归，转化成线性相关

**2.1.4多重共线性**

多重共线性指的是，解释变量之间存在近似的线性相关，一般认定回归系数在0.8左右或者以上可以认定为多重共线性，若出现多重共线性的情况，需要保留比较重要的因素，删去次要的因素保证模型的正确性。

**2.2 损失函数最小值**

一般的方法就是最小二乘法，梯度下降法。

**2.2.1最小二乘法**

最小二乘法又称为最小平方法，是一种数学优化技术，它通过最小化误差的平方和寻找数据的最佳函数匹配。利用最小二乘法可以简便的求得未知的数据，并使得这些数据与实际的数据之间误差的平方和为最小。用最小二乘法来处理一组数据时，可以从一组测定的数据中寻求变量之间的依赖关系，这种函数关系称为经验公式。最小二乘法还可用于曲线拟合，其他一些优化问题也可通过最小化能量或最大化熵用最小二乘法来表示，适用于代数领域。最小二乘法又是一种在误差估计、不确定度、系统辨识与预测等数据处理诸多学科领域得到广泛应用。一般公式是:a=y(平均)-b\*x(平均),本文中未用到此方法，只是介绍有两种求损失函数最小值的方法，所以不进行更多介绍。

**2.2.2 梯度下降法**

**2.2.2.1梯度**

在大学课程高等数学中曾经学习到梯度的相关概念，对多元函数的参数求偏导数，然后将每一个参数的偏导数用向量的方式表示出来。例如f（x，y）=x+y+1，对应的梯度向量就是，简称gradf(x，y）或者-f（x，y）

那这个梯度有什么意义呢，从它的几何意义来看，就是函数变化量最快的地方，意思就是可以求得按照梯度方向上的最大值，同理也可求得反方向上函数的最小值。

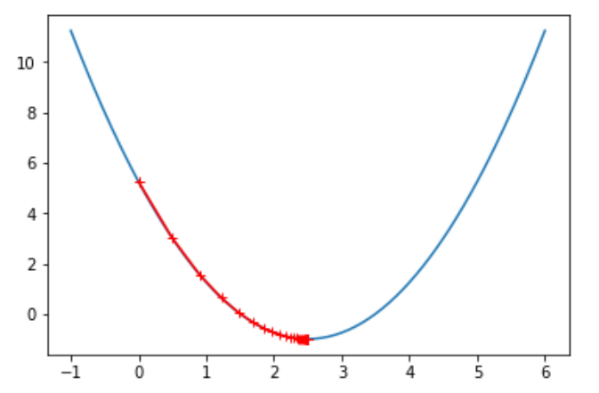
**2.2.2.2梯度下降与梯度上升**

梯度下降法，任务是求得损失函数的最小值，每一步都需要选择坡度变化率最大的方向往下走，每走一步，都需要重新计算函数在当前点的梯度，然后重新选择方向，随着每一步的迭代，梯度不断地减小，到最后减少为零，零是理想条件，一般很难达到，当梯度下降迭代次数足够多的时候，损失函数趋于一个值，这就是损失函数最后的值。

梯度上升发，任务是求得损失函数的最大值，与梯度下降方向相反，于是便称为梯度上升。

**2.2.2.3梯度下降算法详解**

想象你伸手不见五指的黑夜身处一座地形复杂的山中。你看不清山的全貌，只能根据重力感知立足之处的坡度。如果想要下山，你该怎么做？你根据此处的重力感觉，向下坡最陡的方向迈一步。然后根据新位置的坡度，再向下坡最陡的方向迈出下一步。如果来到一个位置，你感觉已经站在了平地上，再没有下坡的方向，你也就无从迈出下一步了。此时你可能已经成功到达海拔最低的山底，但也有可能身处半山腰处一块平地或者一个盆地底部。

****这就是梯度下降法的形象描述。山就是三维空间中函数的图形。某个位置下坡最陡的前进方向就是函数在该点梯度的反方向。向前迈一步，就是让自变量沿着梯度反方向运动一个距离（步长）。感觉已站在平地，就是当前点的梯度为 0，如图1所示。

**图1 梯度下降法示意图**

**2.2.2.4批量梯度下降，随机梯度下降，小批量梯度下降**

（1）批量梯度下降

批量梯度下降是最原始的形式，它是指在每一次迭代之后使用所有样本来进行梯度上的更新，最小化损失函数，使得到的是全局最优解。

（2）随机梯度下降

由于批量梯度下降每一次更新需要对整个样本进行更新，如果数据量很大，工作会很复杂，随机梯度下降就是为了减少工作量，只选取一个样本进行更新。

（3）小批量梯度下降

为了协调稳定性和速度，小批量梯度下降应运而生。其实很简单，小批量梯度下降法和前面两种梯度下降的主要区别就是每一步迭代过程中目标函数的选择不同。小批量梯度下降是从m个样本中随机且不重复地选择n个损失函数的求和。

（3）

**2.4数学知识基础**

**2.4.1期望**

在概率论和统计学中，数学期望是实验中每次可能结果的概率乘以其结果的总和，是最基本的数学特征之一，反映随机变量平均取值的大小，当重复次数接近无限大，数值的算数平均值一定收敛于期望值，用E（x）表示。

**2.4.2方差与标准差**

在概率论和统计中，方差是衡量随机变量或一组数据时离散程度的度量。方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。样本方差是各个样本数据和平均数之差的平方和的平均数。在许多实际问题中，研究方差即偏离程度有着重要意义。

方差的公式：D（x）=E{} (4)

标准差就是方差的平方根。

**2.4.3归一化**

当数据大小差距过大的时候，可以将数据映射到（0,1）之中，一种方法就是把有量纲表达式变为无量纲表达式。主要是为了数据处理方便提出来的。有很多种归一化的方法，在本文中所采用的公式是:

x=(x-x的期望)/x的标准差 （5）

x就是数据集中的解释变量,可能并不是单一数据，一般是一个矩阵。

**2.4.4损失函数**

损失函数是一个用来估计模型的好坏程度的一个非负整数函数，当损失函数很小时，说明模型与目标函数拟合程度很高，反之就说明模型拟合程度很差，损失函数可以根据分类的预测结果，衡量出模型预测能力的好坏。损失函数一般有，0-1损失函数，平方损失函数，绝对值损失函数，对数损失函数，hingleloss。

**2.4.5代价函数**

损失函数是在单个样本上的损失值，而代价函数是相对于整个训练集的误差，相当于损失函数的平均值。常见的代价函数有：均方误差，均方根误差，平均绝对误差，交叉熵代价函数。均方误差，均方根误差和平均绝对误差都可用于回归问题。

**2.5失业率与GDP**

**2.5.1失业率**

失业率指的是失业人口占劳动人口的百分比，失业率在金融市场上占有举足轻重的地位，被称为经济指标上的“皇冠”，1962年经济学家提出“奥肯定律”，揭示了失业率和经济增长的关系，失业率降低说明经济状况良好，有利于货币升值，失业率增长说明，经济发展缓慢，国家会通过政策刺激经济提供更多的就业机会，一般来说发达国家都会把失业率控制在4%到8%，因为部分城市人口缺乏登记动力，而农村人口又不计算在内，所以失业率与真实失业率会有所偏差。

**2.5.2 GDP**

GDP,全称是Gross Domestic Product，中文名称是国内生产总值，是按市场价格计算的一个国家或地区所有的常住单位在一定时期内生产活动的最终成果，中国一般采用生产法计算GDP，计算公式为GDP=产出总额-中间消耗，缺点在于统计相对不够全面，往往忽视了金融业，服务业等非生产部门对于经济的辅助贡献，但是优点在于能够有效的避免重复计算，更重视社会产品的价值，更加适用于中国当下的经济发展阶段。

**2.6监督学习和无监督学习**

在机器学习中，一般分为监督学习和无监督学习，还有半监督学习，主要介绍一下监督学习和无监督学习。

**2.6.1监督学习**

监督学习是指给目标算法一个数据集，里面包括了正确答案，例如预测房价问题上面，给它一个房价数据集，在这个数据集的每个样本中都给出正确的价格，即这个房子的实际卖价，算法的目的就是给出更多正确答案，然后通过算法大致预测一下某个条件下的房价是多少。本文中所采用的是通过往年的考研相关资料以及失业率,GDP来完成分析与预测，就是一种监督学习。

**2.6.2无监督学习**

无监督学习可以把指定的数据集分成几个不同的簇，这就是聚类算法，实际上它被运用到很多地方，比如搜狗的浏览器上面，当点开某个链接，旁边就会有很多类似的链接弹出，搜索一条信息，会有成千上万条信息放在一起。所以无监督学习其实就是给出算法大量的数据，要求它找到自己的分类。无监督学习应用十分广泛，因为往往在现实生活中，难以对物品进行分类，或者分类的成本十分昂贵，无监督学习就应运而生。

**3基于回归模型对考研人数增长原因分析的设计**

一般的分析步骤是，变量因素的选择，建立模型，得出函数，填入数据预测对比。

**3.1变量因素的选择**

根据在第一章中的目的意义背景，本文选取考研人数（Y）被四个主要因素影响，分别是，录取人数，代表了国家对于考研的政策；失业率，代表了社会寻求工作给予学生的压力；GDP，代表社会高速发展带来的收益；毕业生人数，代表了学生之间的竞争压力；有了压力才有动力。

**3.2建立模型**

搜寻数据集，通过数据集分为自变量和因变量，将自变量和因变量的图像对应画出来，分析图像是否满足线性相关，若非线性相关可以将其转换为线性相关，然后建模。

**3.3降低损失值，提高模型拟合度**

通过现有的一些损失值降低的方法，例如最小二乘法，梯度下降法，对于损失值进行降低，并得到相关系数以及迭代之后的损失值，验证损失值的大小，并考虑回归系数的大小，得出结论。将数据代入回归方程，与原数据对比，观测模型拟合度。

**4实验结果与分析**

**4.1确定模型**

数据来源：

近十年来的报考人数和录取人数:

<https://cdn.hubblecontent.osi.office.net/icons/publish/icons_sun/sun.svg>

近十年的失业率:

<https://www.chyxx.com/industry/202005/866456.html>

近十年的gdp：[https://www.kylc.com/stats/global/yearly\_per\_country/g\_gdp/chn.ht](C:\\Users\\ZHURUIHONG\\Pictures\\5ae2cecf00015eec06000458.jpg)

[ml](C:\\Users\\ZHURUIHONG\\Pictures\\5ae2cecf00015eec06000458.jpg)

近十年的大学生的数量:

<http://www.360doc.com/content/20/0707/16/37637738_922802506.shtml>

根据给定的数据可以将数据列出来更加直观的（2000-2019），如表1。

“Enrollment”:研究生录取人数。

“u\_rate”:失业率。

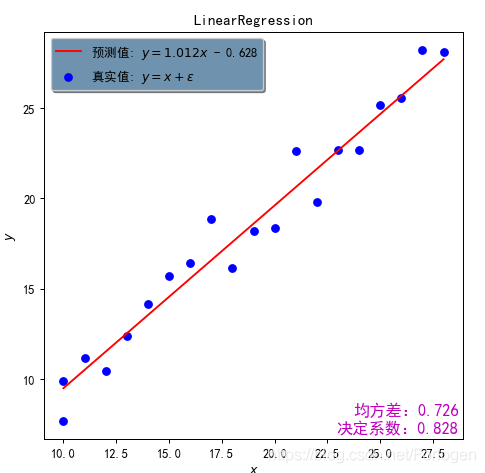
“GDP”:国内生产总值。

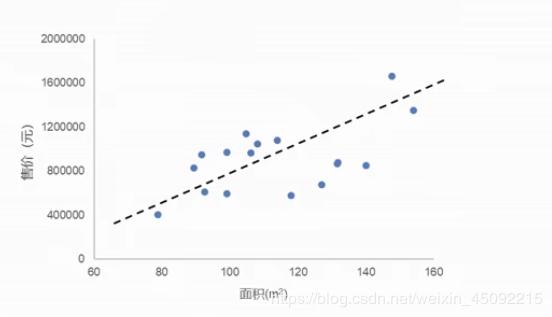
“graduate”：大学生毕业人数（面临毕业的大学生人数）。

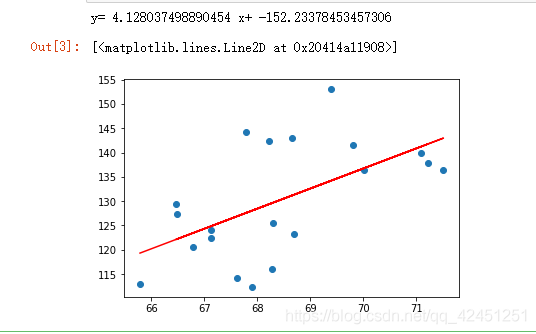
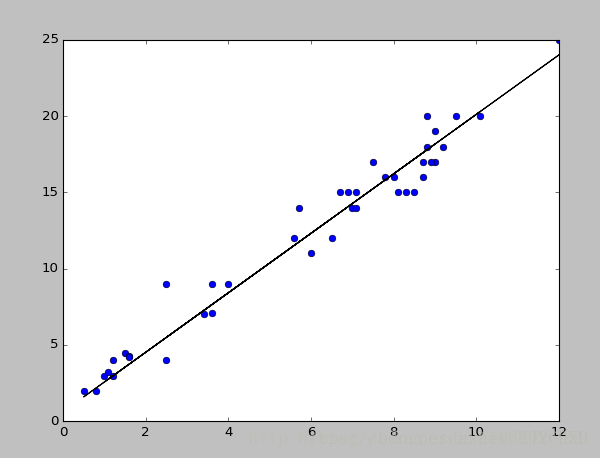
“applicant”：研究生报名人数。

**表1 因变量数据**

根据所给的数据，将毕业人数，录取人数，GDP,失业率各自作为单独自变量，以考研报考人数作为因变量，通过python所给的matplotlib包进行绘图得到下面四张图，如图2-图5所示。







通过分析以上四张图，可以大致得到，考研录取人数，失业率，GDP，毕业生人数是与考研报名人数存在线性关系的，于是可以采用多元线性回归模型进行分析预测。

**4.2 计算**

**4.2.1数据预处理**

先将数据进行预处理，方便后续使用。将之前所得到的数据做切片处理，将前四部分（考研录取人数，失业率，毕业生人数，GDP）作为自变量，最后一个部分（考研报考人数）作为因变量，由于数据的大小量是不一样的，例如失业率只有4%，报名人数就达到了290万甚至更高，这样会影响到后面数据分析的结果，所以在这里进行一步归一化处理。

这样就可以得到以下数据，如表2，表3。

**表2 解释变量与时间**

**表3 被解释变量与时间**

将x，y化为矩阵方便下一步的计算.

数据就处理完毕了，有些数据跳出了[-1,1]，这个其实影响不大，在[-3，3]的范围之内。

**4.2.2通过梯度下降计算**

由于有四个自变量，一个因变量，于是可以设函数

y=+（e表示error）

\*——=J（θ） （6）

J（θ）=

（7）

于是可以把公式写成(总共有4个因变量，20条数据):

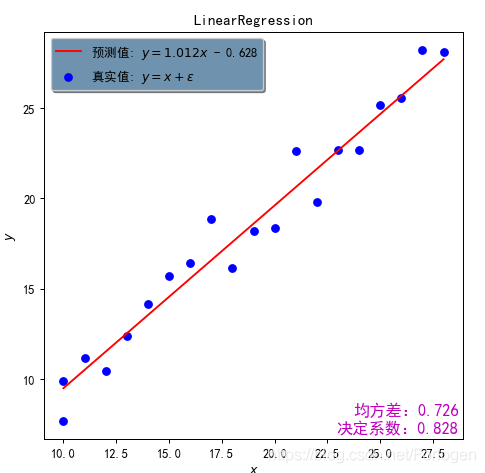
\*—=J（θ） （8）

所以需要做的事情就是求出j（θ ）的最小值，让损失函数最小化，这里就要用到梯度下降，让J（θ）对，，，求偏导

这里就要强调一下批量梯度下降的优点了，可以全局同步更新，求得全局最优解。

Temp0=-

Temp1=-

****Temp2=-

Temp3=-

=Temp0

=Temp1

=Temp2

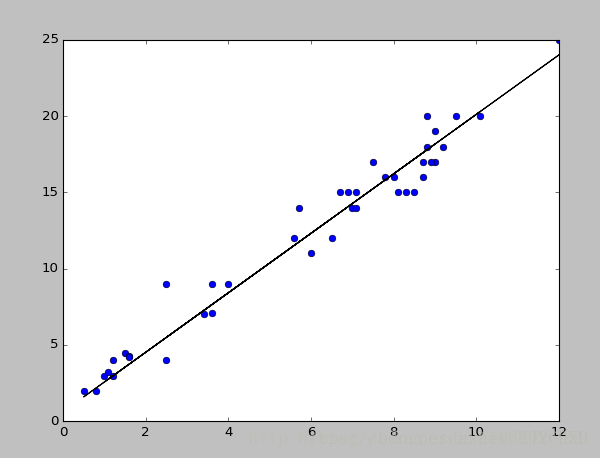
=Temp3

代码部分如图6

**图6 梯度下降函数**

x和y分别是数据处理后转成矩阵的数据，alpha是学习率，theta是θ，也就是回归系数，

iters是迭代次数，如图7还有一个计算cost的函数：



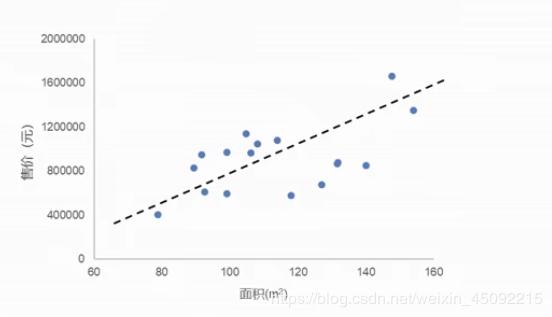
**图7 cost函数**

代码部分就是将公式转换成代码，并不复杂。

输入的是alpha和iters，输出的是θ和cost。

通过本公式可以计算出迭代之后的θ和损失值。

如图8可以直观看出迭代次数和损失值关系

**图8 迭代次数与损失值**

随着迭代次数（共1000次）的增加，损失值趋于一个值，通过程序算出来

经过迭代之后的cost的值是:0.05721332082432653

计算出相关系数:

相关系数是:[[ 0.30134074 ，-0.00536811 ， 0.315357 ，0.29089532]]

于是得到回归方程:

Y=0.30134074-0.00536811+0.315357+0.29089532+0.05721332082432653

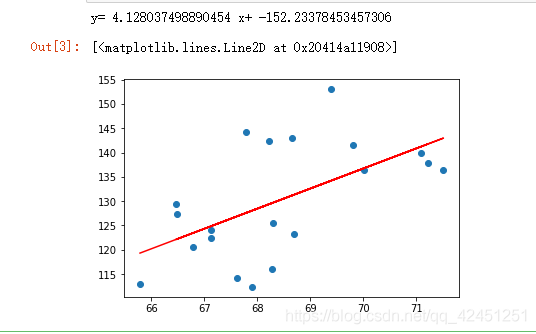
在回归方程中，回归系数代表影响程度，回归系数越大，影响程度越大。正回归系数表示x增大也y随着增大，负回归系数表示x增大，y反而减少。X1代表录取人数，X2代表失业率，X3代表GDP，X4代表在校大学生人数，y代表考研人数。

由该模型可知，考研录取人数每增加一万人，考研人数就会增加0.30134074万人；失业率每下降一个百分比，考研人数就会减少0.00536811万个；GDP每上升一万亿，考研人数就会增加0.315357个；在校大学生每增加一万人，考研人数就增加0.29089532万人。

从这个回归方程中能获取到的信息，考研人数的增长与社会发展息息相关，社会发展迅速，不仅仅提高了人们的生活整体质量，还增加了人们对于高学历的需求，同时考研人数的增长也与国家政策离不开关系，国家政策的不断改革，大力支持研究生的政策，在某种程度上反映了国家对于高学历人才的需求，其次就是高校毕业生对于考研人数的影响，高校毕业生的增长，导致了本科生的竞争能力较弱，本科生的竞争压力增大，无疑导致了不少人选择考研提高自己的竞争能力，失业率对于考研人数是负相关的，在某种程度上反映出来考研人数的增多，能够减缓就业压力，也会反向促进考研人数的增长。

**4.2.3预测值对比**

通过sklearn中的Linear\_model的LinearRegression，将数据分为测试集和训练集，利用二十个数据按照训练集：测试集=7：3，训练出回归模型，然后对于2020年的数据进行分析，2020年新增加疫情，疫情在某种程度上影响了考研人数，例如原定出国的同学，由于国外疫情严重，选择了考研，这些人在名牌大学中占比很大。由于不好量化，所以将预测的结果与2020年的数据进行对比，这样就大致推测其他因素以及疫情对于考研人数的影响。



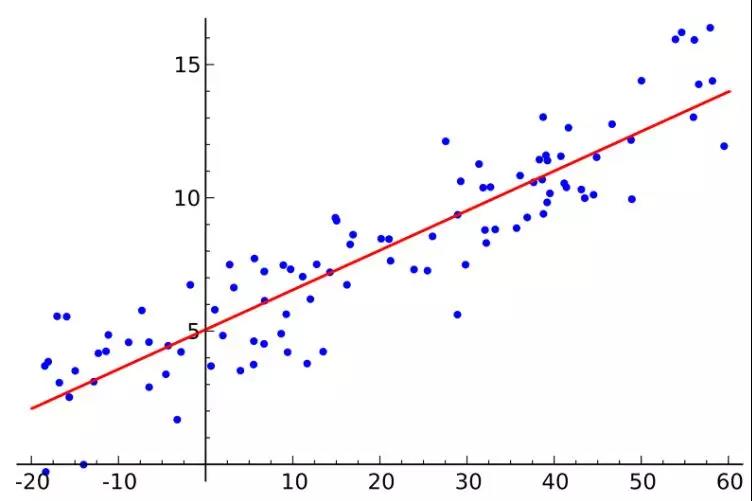
代码部分如**图9：**

图9 预测的另一个算法代码

得出结果，这个模型的分数是:0.8733872630929872

2020年考研人数:[[297.95521985]]。

再使用比较经典的GBDT来预测一下数据，做一下对比，主要代码如图10。

****

**图10 GBDT预测主要代码**

得出结果154.280586，预测结果差距比较大。

利用回归模型预测出来的值是319.375479174拟合程度比这两个算法高，但是也存在其他重要影响因素（疫情等）的问题,导致与实际的377万有差距。

**4.3结果分析**

通过上述分析，可以得知，考研人数的快速增长在某种程度上是一种被动盲目的行为，如果要改变这一点，需要很多方面的一起努力，为了引导考研者做出理性的选择，总共给出如下几点建议。

①政府方面，需要继续优化毕业生就业政策，把毕业生的就业环境变的更

好，同时应在法律法规上加强对于求职者的合法权益的保护，在整体上把握人才培养的规则，给社会输送不同层次，不同领域的人才

②企业方面，企业应树立正确的用人观念，不应固有“唯学校”，“唯学

历”的思想，用人单位需要建立合适的用人观念，在选拔人才时，用人的成本计算时应采取理性的良性标准，以免错失了一些能力较强的人才

③学校方面，学校应多多开展对于学生职业规划的活动，同时要加强学生

的实践动手能力，切不“纸上谈兵”,提高学生的综合素质能力，实现学校与社会的接轨，防止学生只学到书本上的知识，动起手来却不知何从下手。

④考研者自身方面，要确立自己的目标，知道自己真正需要的是什么，制

定适合自己的职业生涯规划。

**4.4本章小结**

通过往年的数据来分析未来的数据，是一种监督学习的典型案例，区别不同就在于，近些年来，没有人做过利用编程软件在考研人数原因上的分析和预测，

基本上都是利用现有的Eviews和spss等数据分析的软件进行分析，其中的算法是已经既定好的，如果有更好的算法被发现，就需要去修改底层原理，这是现有的软件无法做到的，所以通过编程，以及对于算法的熟悉，对于以后优化算法有一定的帮助，同时在学习的同时，学到了很多统计学，经济学的一般统计方式，对数据分析的模式，对于计算机行业来也是很有帮助的，分析到数据背后的隐藏含义，这也是以后需要做的工作之一。

**5 总结**

**5.1本文算法的优点**

当用梯度下降与多元线性回归模型结合时，当迭代次数达到一个数时，损失值趋于一个值，利用到全部样本，更新的时候也是全局更新，回归系数每次更新，损失值可以降到很低，拟合程度相对而言比较高。

**5.2本文的缺点和未来研究展望**

首先是对于考研这个问题，其实笔者开始本来是想做关于每个院校的每个专业的考研分数的预测。因为很多同学跟笔者有过一样的困扰，很多院校的某些专业，假如今年复试线录取分数很低，那么明年就会很多人扎堆报名这个院校，水涨船高，导致分数线上涨，来年大家看到分数线很高又不敢报名，又会导致这个学校分数线变低，循环往复，就会导致考生对于自己的不自信，可能往往会选取低于自己目标的院校。笔者想做的就是把往年的数据都收集起来，根据报录比等一些比较重要的原因，对数据进行预测（就跟景区预测人数一样），但是问题在于当去搜寻很多学校的时候，数据缺失很严重，往往只有3-4年的数据，数据集的缺失，导致进行不了下一步的工作，只能无奈放弃。希望以后能够保留这一些数据，可以进行一部分的预测，给学弟学妹们一点小建议，给他们划分大致的区间，让他们更加明确自己所要报考的院校。

其次，就是对于这个项目，自变量可能会互相影响，即多重共线性，因为可能就GDP与失业率有关系，就忽略了这一点，但是这一点肯定是对预测结果的准确性有影响的，还有很多其他的因素都会影响到最后的预测结果，例如2017年的时候考研录取人数达到了80.5万人，而2016年只有51.72，2018年也只有72.2万人，是因为2017年的时候统考第一次加上了在职，因为这种可能还有其他的因素是没有考虑在内的，这些也会导致最后的预测结果的不准确。

然后就是这个算法的问题，当使用梯度下降方法对模型参数，需要遍历所有的训练数据，当迭代次数很高时，往往需要很大的计算量，需要花费很长的时间，

在实际应用中基本上不可行。而且由于开始的时候看到他们是满足大致的线性相关，就做了线性回归模型的学习，但是应该有更好的模型能对于这个东西的预测，由于知识的匮乏导致只能用到这个回归模型，研究生期间学完机器学习和人工智能，对数据结构算法进行更深一步的学习，可以优化一下这个模型，能得到更加准确的结果。

最后，这个项目是可以扩展的，目前能够实现的是关于全国的研究生考研因素的影响，可以将全国拓展延伸到每个院校，针对每个学校的改革政策，分析出其中影响较大的因素，用更加适合的模型，更加合适的算法，算出精确度更高的回归模型，能针对每个院校的考研人数增长的原因给出具体的建议，而不像现在，只能给全国的考研人数增长的现象给出笼统的意见。

**5.3个人总结**

通过对于机器学习的初步学习，让笔者有了更加浓厚的兴趣，对于未知的知识的渴求，同时也是十分的好奇python自带的那些用于机器学习算法的底层原理，想知道是如何实现的，在慢慢的熟悉的过程中，了解到自己的不足之处，对于一些高等数学和线性代数里面的基础知识的薄弱，让笔者无法理解一些计算过程，在研究生前这一段时间，需要多花时间补一补线性代数和高等数学，还有一些数据结构的内容，以免在研究生期间拖后腿。