

**基于python的数据挖掘**

**----阳光集团的具体数据挖掘项目**



2018-7-24

江苏阳光集团

江苏省江阴市

作者签名

**基于python的数据挖掘**

**----阳光集团的具体数据挖掘项目**

**摘要：**随着大数据时代的到来，商业、政府机构等许多部门和行业都面临着数量越来越多的数据资料，因此数据挖掘受到外界越来越广泛的关注。本文就阳光集团所提供的广西省的具体数据，基于python语言对其进行数据清洗、数据挖掘以及回归分析。

**关键字：**python、数据挖掘、数据清洗、回归分析

## 1.涉及技术简介与具体项目背景

[Python](https://en.wikipedia.org/wiki/Python)是一种面向对象的解释型计算机程序设计语言，由荷兰人Guido van Rossum于1989年发明，第一个公开发行版发行于1991年。Python语法简洁清晰、封装性很好并且拥有丰富且强大的库。

[数据挖掘](https://en.wikipedia.org/wiki/Data_mining)指的是从数量极其广大的数据中通过各种特定算法来搜索其中隐藏着的信息的过程，所以又名为资料勘探。数据挖掘属于计算机科学的一种，但其中综合了多个学科例如统计学、信息检索等方面的内容。

[数据清洗](https://en.wikipedia.org/wiki/Data_cleansing)，顾名思义就是指将“脏”数据清洗干净以符合操作规范。数据清洗的工作内容一般包括一致性检查以及无效值和缺失值的处理。

[回归分析](https://en.wikipedia.org/wiki/Regression_analysis)是回归方程的一项重要应用。所谓预测就是对给定的X值，估计Y值将落在什么范围。回归预测方法有多种类型。依据相关关系中自变量的个数不同分类，可分为一元回归分析预测法和多元回归分析预测法。

本次项目的背景是使用阳光集团所提供的广西省的具体量体数据来进行预测。此数据集中包括了性别、身高、体重、衣长等对回归预测具有重大参考意义的数据。所以，根据这些数据，进行对应的预测就成为了本次项目的目的。例如：当测试一个从未测试过的员工的量体数据时，可以依据他/她的身高、体重数据，来大致预测出该员工其他项数据比如衣长、肩宽等的可能取值。本次数据目的是根据身高、体重的值来预测紧胸围和紧臀围。

## 2.项目所面临的问题以及解决方法

### 问题一：清洗数据（异常值处理、处理方法优化）

1.格式处理

本次项目采用python中的pandas包来进行数据的读取，性别属性的两个属性值“男”、“女”也不便于在pandas中进行操作，因此更换采用0-1型数据格式（“男”为1，“女”为0）。

2.缺失值处理

随后，通过np.where查询语句，我们发现本数据集内有大量值为空（NaN）的属性格，这些数据必定会对数据的处理和预测产生负面的影响。



图1

考虑到本次项目的准确性以及其他数据依旧具有相当的代表性，我们决定将这些数据全部删除。原csv文件中有5998条数据，相关数据经过删除NaN值后，还剩下3292条数据。



图2

3.异常值处理

首先导入matplotlib相关包，然后先对这些输出画出一个箱型图。

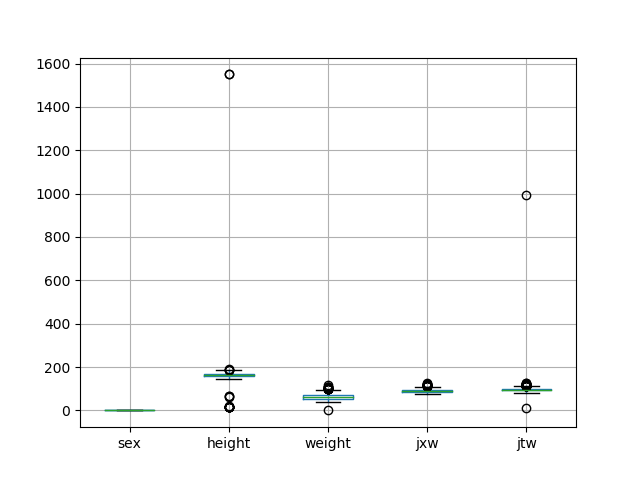


图3

从图3中可以看出数据最开始的离散程度还是很大的，尤其是身高和紧臀围属性，经计算的标准差高达36.049和16.831。通过常理可知，身高和紧臀围1000多的数据必然出现问题，而且有极大可能是厘米和米单位换算出现问题可以暂时当作训练集保留，通过后期的预测来校验；同时明显小于常理值的数据也肯定出现了问题，这些数据大概率是录入数据的时候出现了问题，因此删除。

同时3σ原则也可以很有效的过滤数据。正常的一组数据，mean+3σ到mean-3σ之间就会涵盖总数据的99.7%，所以当将数据集中的数据范围框定在mean+3σ到mean-3σ之间即可进行相对有效的数据过滤。未经过箱型图法过滤的数据过滤后，淘汰的数据恰恰也是1000多的数据和远远小于常理值的数据。但这并不代表箱型图法和3σ原则效果一样。

当数据的离散程度较大时，3σ原则处理数据的效果会由于箱型图法。若一组数据较为集中，则其标准差σ就会非常小，因此淘汰的数据就会相应减少；而箱型图主要是按照percentile四分位值来构图的，当数据的离散程度较小时，四分位差就会显得很小，从而会错过滤很多正常值。然而这次数据是比较集中的人体数据，因此采用3σ原则会优于箱型图法。

在实际使用3σ原则进行处理时，会发现存在部分极端异常错误数值大幅影响标准差的情况。在测试数据中，女性样本存在极端数据1550，由于样本量仅有1000余个，导致这一个极端数据将影响方差致使其增加接近上百，最后标准差的数值增加几十，导致被筛选的数据范围过大。如图4：

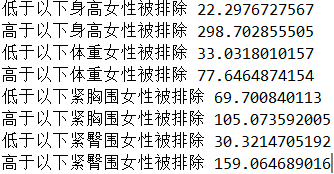


图4

如上图所示，对女性数据进行一次3σ原则筛选后，我们发现身高低于22cm，高于298cm的数据才会被排除，但实际上，存在一些异常数据仍未被排除。例如，部分数据由于录入时的失误，将身高160输入为60，这样的数据仍然被保留了下来，这是因为极端数据身高1550导致标准差σ被拉得过大，筛选范围过宽。

由于存在这种情况，因此有必要进行两次3σ原则筛选，第一次将极端异常的错误数据排除，让标准差相对正常。第二次在标准差较为正常的情况下，继续排除异常数据，保证得到的数据是真正可信的。

对女性数据进行两次3σ原则筛选后，筛选范围如图5：



图5

我们看到，女性身高、紧臀围两项数据的筛选范围有了明显的缩小，而体重、紧胸围没有，这是因为女性身高和紧臀围分别有1550、990这两个极端大的异常数据，大幅影响了标准差，而体重和紧胸围没有偏差如此离谱的错误录入情况。

### 问题二：填写数据错误（反预测）

正如上文所提到的那样，在填写数据时，难免会发生错误。暂且把这些错误分为两种，一种是不可猜测的错误数据，比如将160填写成60或者16等；另一种时刻猜测的错误数据，比如将160填写成1600。实际上，第一种错误因为无法猜测，对于数据的回归和预测是没有正面影响的，然而第二种数据，可以猜测到错误原因的，则需要基于其他数据的预测来检测其正确性。以两组错数据[0,1550,53,84,90]、[0,1550,50,93]为例。除去这两组数据后，再通过对这组数据的梯度下降回归（gradientDescent），经过10000次迭代，可以得到大致的预测结果。当输入[sex, weight]时，可以得到预测结果[height]。带入数据[0, 53]和[0, 50]，分别得到了结果160和159。

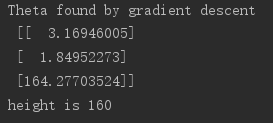
=>

图6

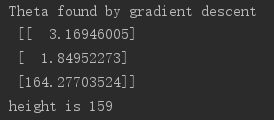
=>

图7

两组数据都在误差范围内，因此可将数据1550改为155填入表内。再测定错误数据[0,158,52,88,994]这组数据中，紧臀围数据994明显高于常理值，因此再带入预测。设定程序为输入[sex, weight]得到[jtw]的预测值，输入数据[0,52]可以得到结果[92]。因此可以相信此条错误数据也是因为单位出现了问题。将99或者94带入即可。

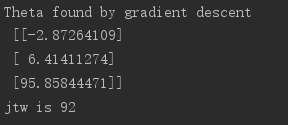
=>

图8

涉及到体重时，考虑到常用单位为斤和公斤，也可能会存在类似误差。第一次3σ筛选后得出的错误数据如图9：

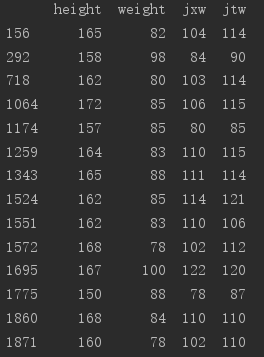
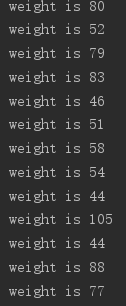
=>

图9

经过预测后，我们发现，体重中有些数值接近于预测值的两倍。于是我们考虑到在体重录入时，员工可能会将自己的体重单位默认为“斤”和“公斤”，这两个单位见正好是两倍关系。所以可以将记录体重约为预测体重两倍的记录认定为单位错误，处理这些数据时候，可以将数据/2，再带入计算。再具体程序中，将1.5倍视为单位错误。通过参数count\_d来记录替换次数。

具体代码如图10：

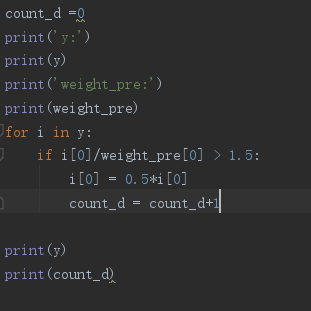


图10

最终得出在所有数据记录中，替换次数为489次。

### 问题三：回归算法比较与优化（均以身高+体重预测紧臀围为例）

这次项目，我们使用了python中的scipy包，它涵盖了很多便捷的数据处理方法，比如回归、聚类、分类等算法。当我们需要对数据进行回归预测时，这个包里的内容给予了我们极大的帮助。但是在诸多回归方法中，应该如何甄别各回归方法的效率呢？我们可以使用scipy包中自带有score的评分包来对回归方法进行评分。

最开始时，我们采取的性别0-1化对项目结果有着超乎预期外的影响，当程序读取整则数据时，（以普通最小二乘法为例）整个算法的预测评分才只有0.65。预测结果如图10和图11。

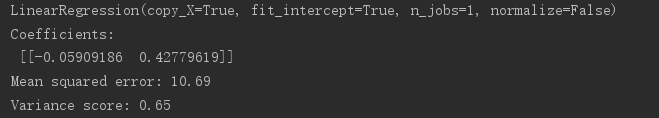


图11

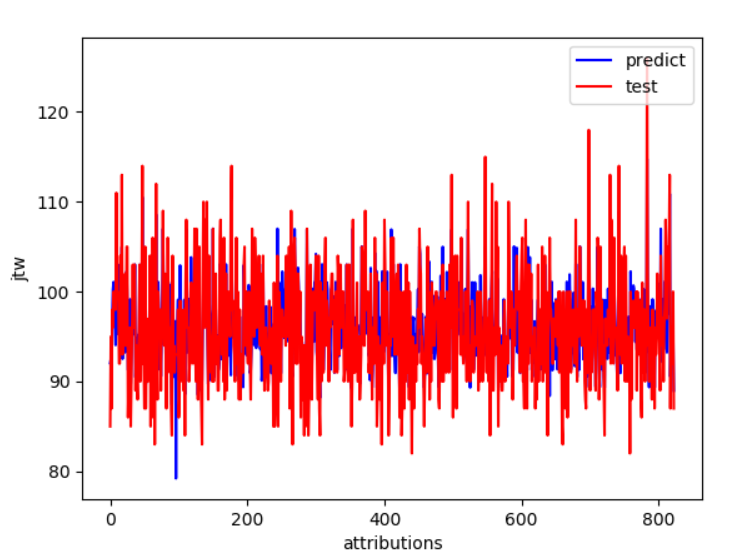


图12

然而当我们将男女分开进行测试时，女生的评分为0.89，男生的评分为0.87。相比较之前合并处理的结果有了很大的提升，对应的预测结果图也显示出了较好的预测结果

女生：

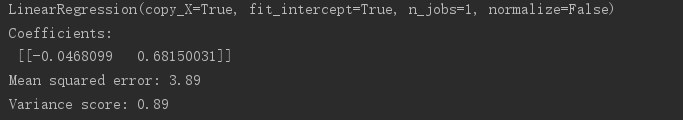


图13

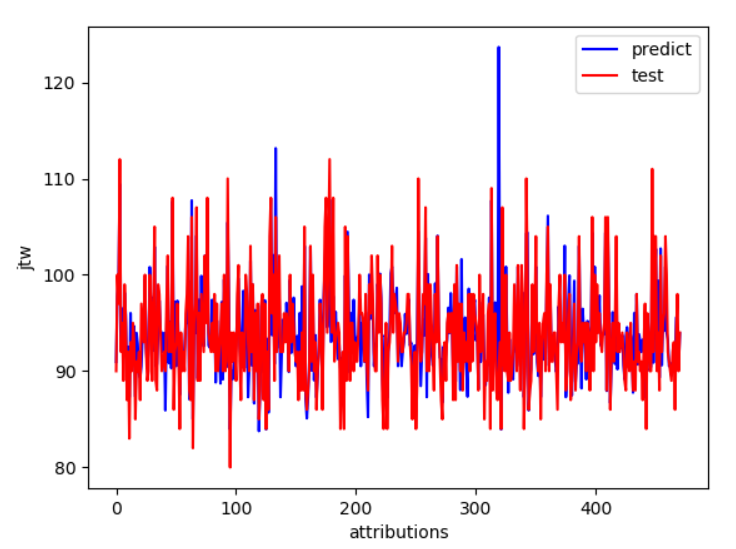


图14

男生：

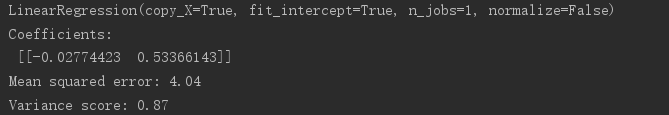


图15

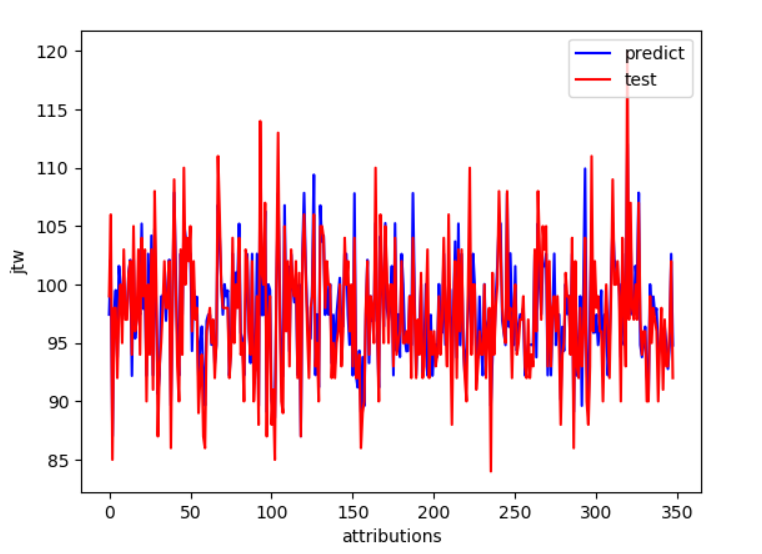


图16

#### [Ordinary Least Squares](https://en.wikipedia.org/wiki/Ordinary_least_squares)普通最小二乘法（0.64，0.66）

普通最小二乘（OLS）是一种类型的线性最小二乘用于估计未知方法参数在一个线性回归模型。OLS 通过最小二乘法原则选择一组解释变量的线性函数的参数：最小化给定数据集中观察到的因变量（被预测变量的值）与预测变量之间差异的平方和通过线性函数。

几何上，这被视为与因变量轴平行的平方距离之和，集合中的每个数据点与回归表面上的对应点之间的差异 - 差异越小，模型越适合数据。得到的估计量可以用简单的公式表示，特别是在右侧的简单线性回归（单个回归量）的情况下。

核心代码如图17：

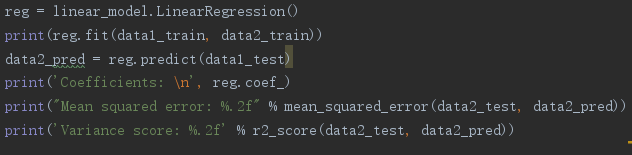


图17

女生：0.64

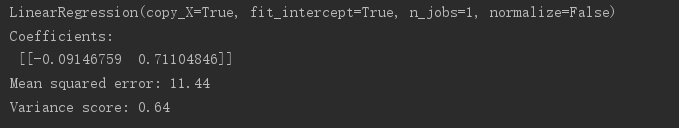


图18

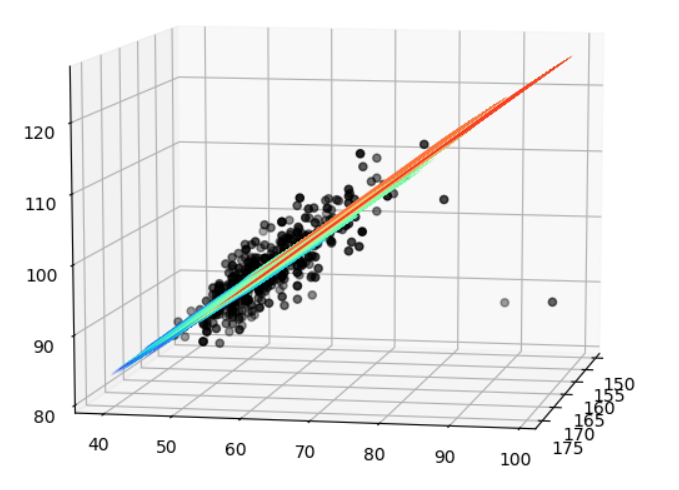


图19

男生：0.66

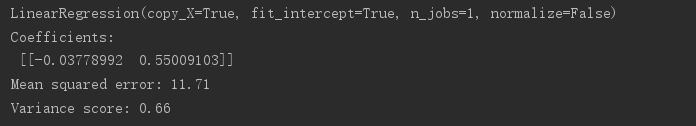


图20

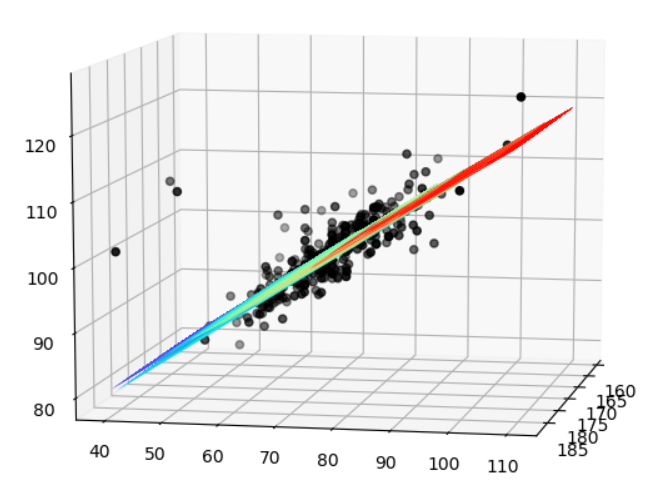


图21

#### [Ridge Regression](https://en.wikipedia.org/wiki/Ordinary_least_squares)岭回归（吉洪诺夫正规化）（0.80，0.78）

岭回归(英文名：ridge regression, Tikhonov regularization)是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。

核心代码如图22：

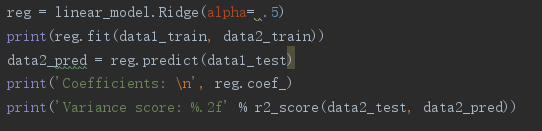


图22

女生：0.80

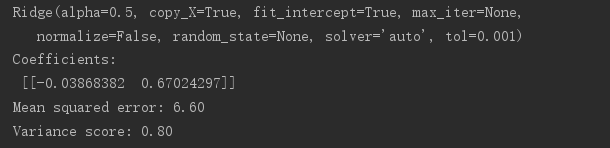


图23

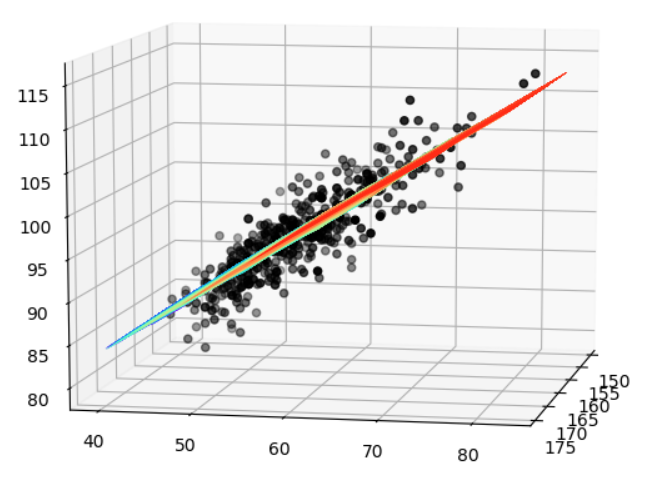


图24

男生：0.78

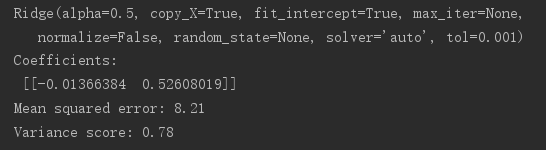


图25

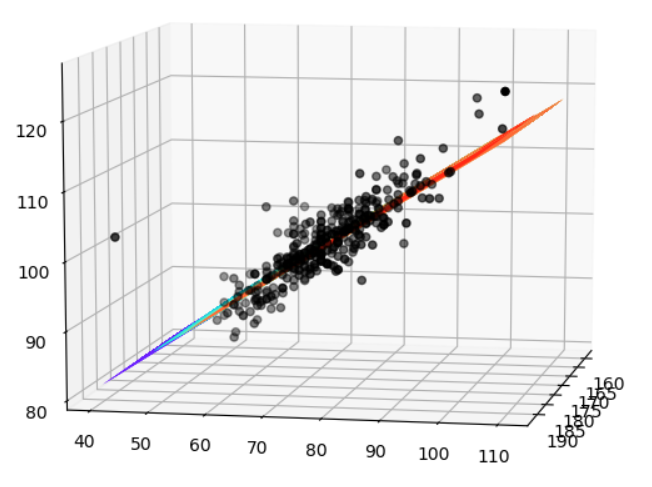


图26

利用RidgeCV函数，可以输入一批alpha值，进行验证（默认为留一交叉验证），获取使得岭回交叉验证值（默认为均方差）平均最低的alpha，使岭回归系数alpha最优化。其中，参数stroe\_cv\_values=True可以保存岭回归的交叉验证值。

为了能够自动化获取有一定精度的alpha值，进行两次RidgeCV函数的alpha最优检索，第一次在10e-10到10e10的范围内按等比数列选取200个alpha值，获取与最优alpha的数量级相同的预计值。第二次在alpha的数量级范围内按一定精度寻找最优alpha值。代码如图27：



图27

得到的男性、女性的紧胸围、紧臀围的预计最优alpha值与最优alpha值如图28。

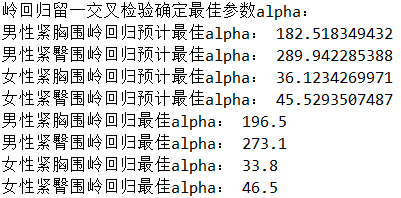


图28

绘制不同alpha下得到的岭回归交叉验证均值图，得见最佳alpha处对应着岭回归交叉验证均值的最小值。代码如图29

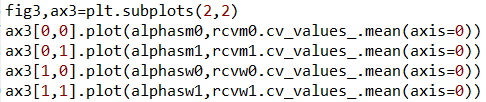


图29

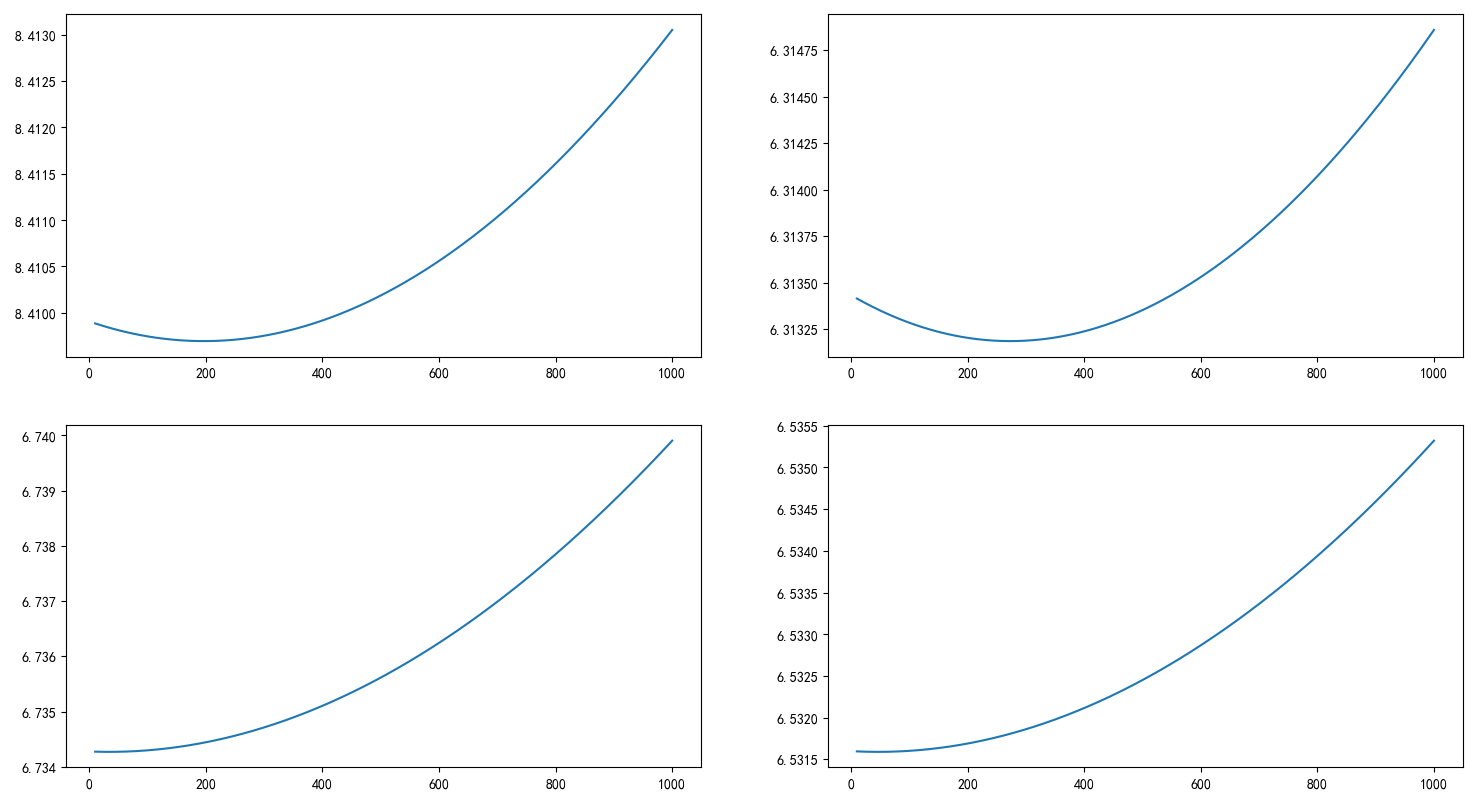


图30

根据得到的最佳alpha系数再次对数据进行岭回归计算。如图31：

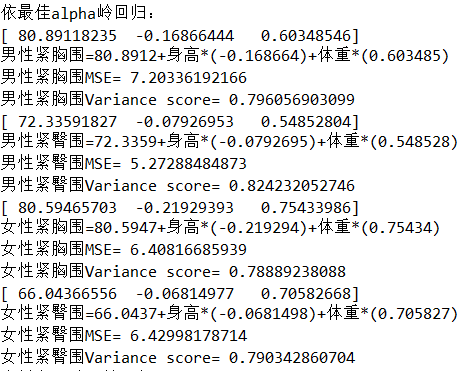


图31

#### [Lasso](https://en.wikipedia.org/wiki/Lasso_(statistics))套索回归（0.74，0.80）

在统计学和机器学习中，套索（最小绝对收缩和选择算子）（也称为Lasso或LASSO）是一种回归分析方法，它执行变量选择和正则化，以提高其产生的统计模型的预测准确性和可解释性。Lasso最初是制定的最小二乘模型，这个案例揭示了估计量行为的大量数据，包括它与岭回归和最佳子集选择的关系以及套索系数估计与所谓的软阈值之间的联系。它还揭示了（如标准线性回归）如果协变量是共线的，系数估计不必是唯一的。

尽管最初定义为最小二乘法，但是套索正则化很容易扩展到各种统计模型，包括广义线性模型，广义估计方程，比例风险模型和M-估计，以直接的方式。Lasso执行子集选择的能力依赖于约束的形式，并且具有各种解释，包括几何，贝叶斯统计和凸分析。

核心代码如图32：

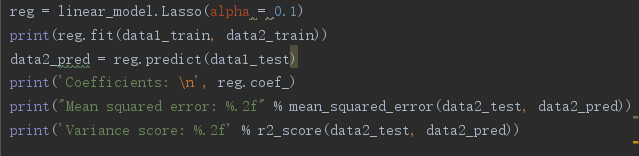


图32

女生：0.74

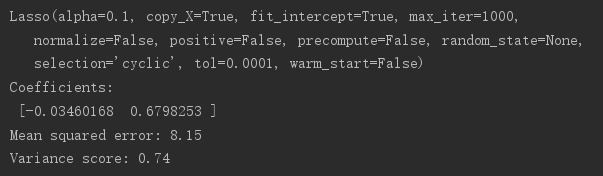


图33

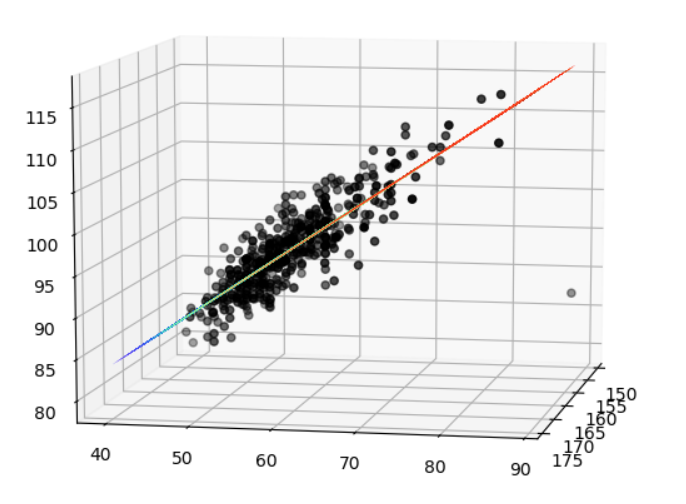


图34

男生：0.80

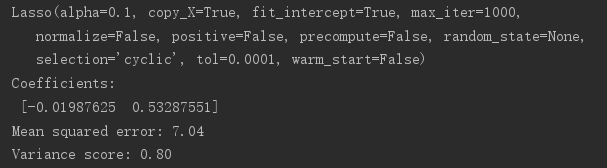
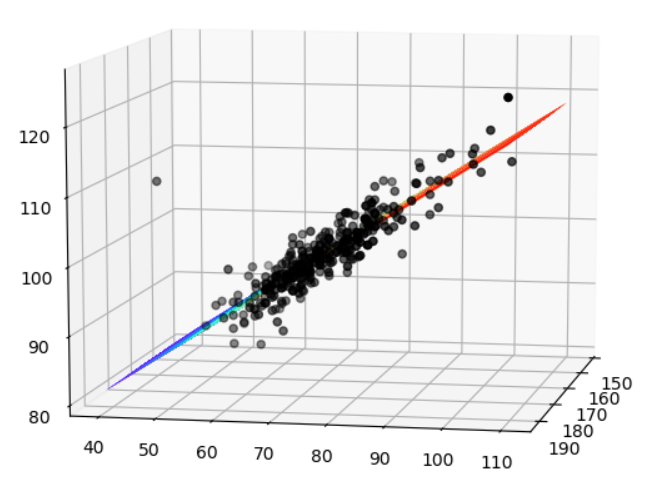


图35

  
图36

#### [LARS](https://en.wikipedia.org/wiki/Least-angle_regression) 最小角Lasso回归（0.30, 0.25不适用）

在统计学中，最小角度回归（LARS）是一种将线性回归模型拟合到高维数据的算法，由Bradley Efron，Trevor Hastie，Iain Johnstone和Robert Tibshirani开发。

LARS解决方案不是给出矢量结果，而是由表示参数矢量的L1范数的每个值的解的曲线组成。该算法类似于前向逐步回归，但不是在每个步骤中包括变量，而是在与残差的每个相关性等角度的方向上增加估计参数。

核心代码如图30：

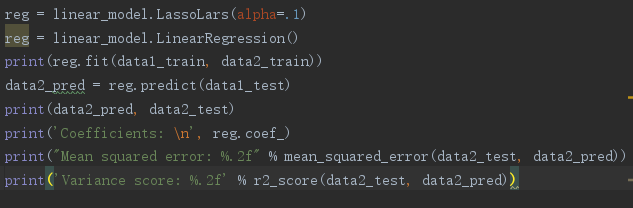


图37

女生：0.30

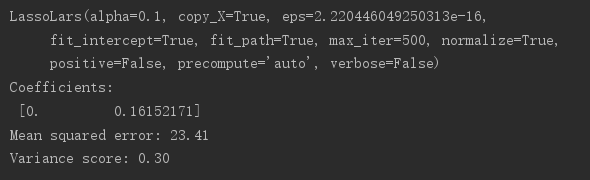
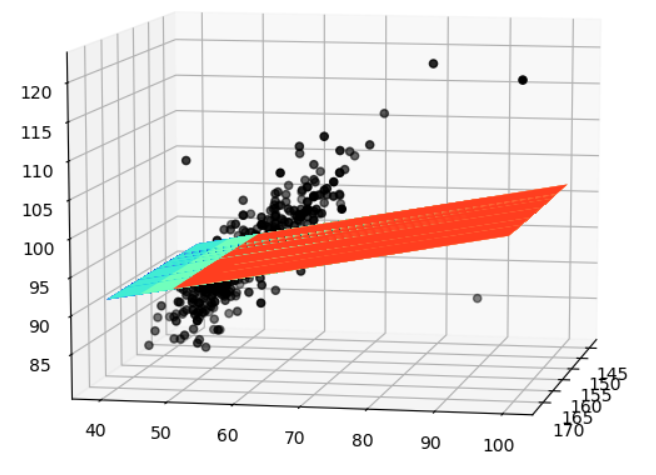


图38

  
图39

男生：0.44

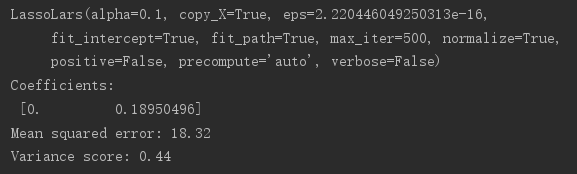


图40

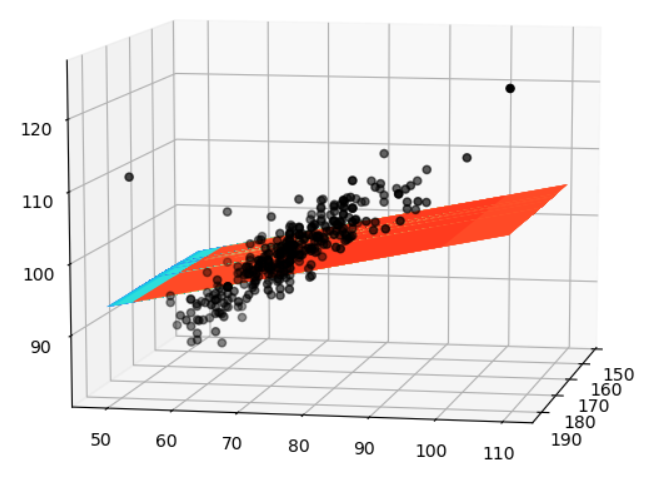


图41

#### [Bayesian Ridge Regression](https://en.wikipedia.org/wiki/Ordinary_least_squares)贝叶斯岭回归（0.72，0.77）

在统计学中，贝叶斯线性回归是一种线性回归方法，其中统计分析是在贝叶斯推断的背景下进行的。当回归模型具有具有正态分布的误差时，并且如果假设特定形式的先验分布，则可以获得模型参数的后验概率分布的显式结果。

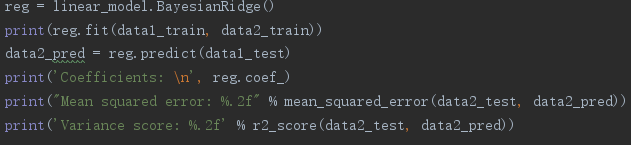


图42

女生：0.72

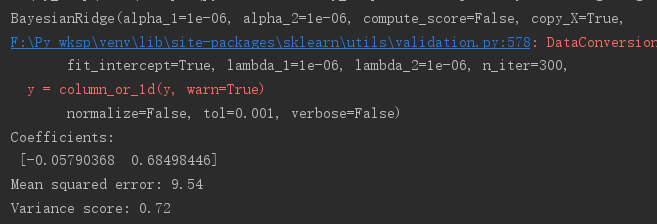


图43

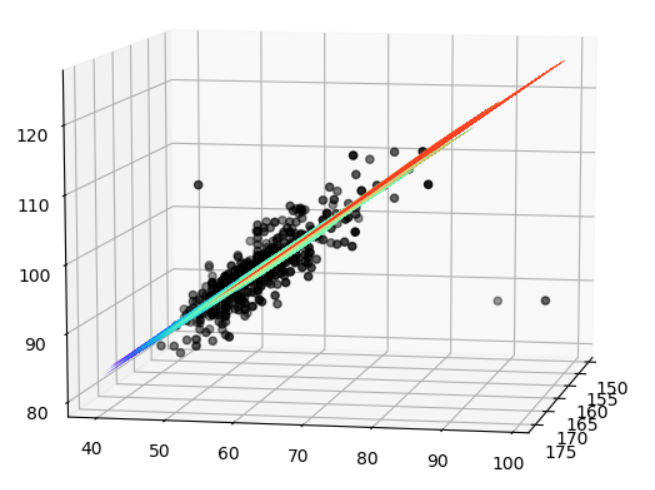


图44

男生：0.77

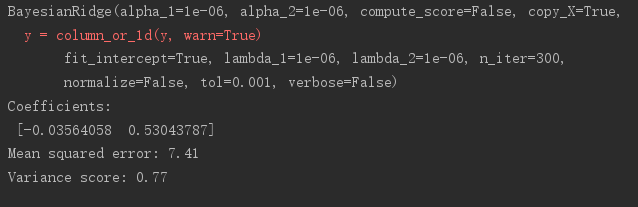


图45

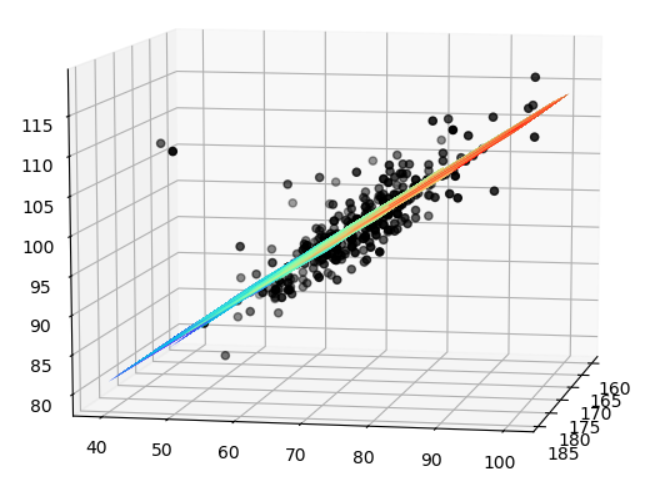


图46

#### [Polynomial regression](https://en.wikipedia.org/wiki/Polynomial_regression)多项式回归（0.74， 0.72/0.78， 0.75）

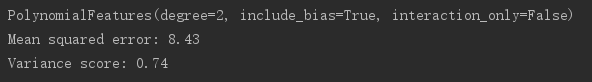
在统计，多项式回归是一种形式的回归分析，其中所述的关系独立变量 X和因变量 ÿ建模为Ñ个程度多项式在X。多项式回归拟合的值之间的非线性关系X和相应的条件均值的ÿ表示为E，（ÿ | X），并已被用于描述非线性现象，例如组织的生长速率，湖泊沉积物中碳同位素的分布，[2]和疾病流行的进展。虽然多项式回归拟合数据的非线性模型，但作为统计估计问题，它是线性的，在某种意义上，回归函数E（y | x）在从数据估计的未知参数中是线性的。因此，多项式回归被认为是多元线性回归的特例。

核心代码如图47：



图47

女生：degree=2对应0.74 degree=3对应0.78



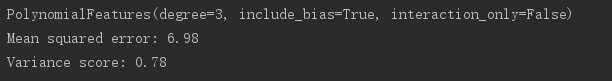
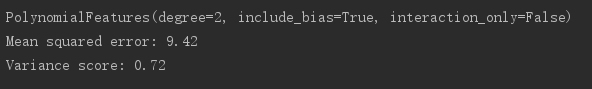


图48

男生：degree=2对应0.72 degree=3对应0.75



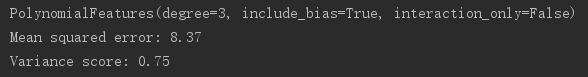


图49

可以发现，当对应指数上升时，对应的算法评分也会上升，准确率也会提高。但是，过高的次数也会带了过拟合化的问题。如图50：

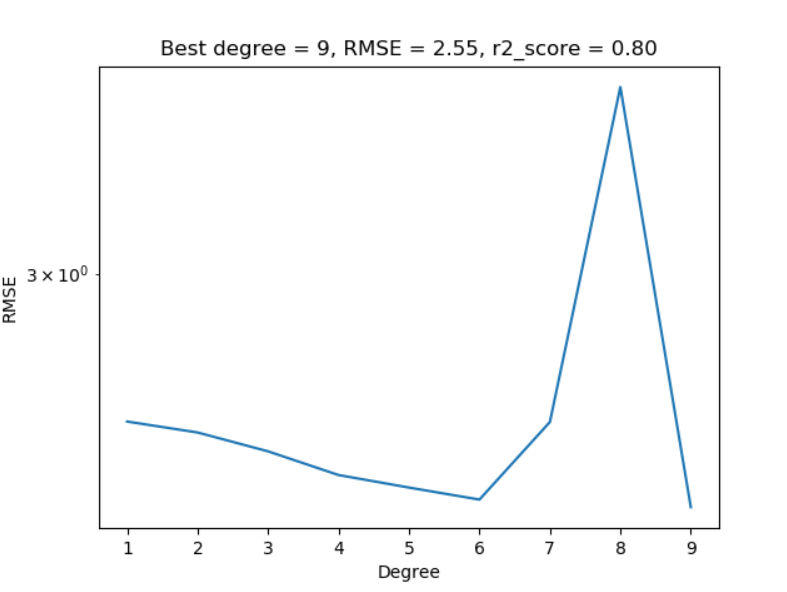


图50

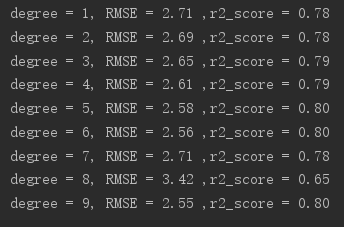


图51

RMSE(Root Mean Square Error) 均方根误差是由一个预测值之间的差异频繁使用的测量模型或估计量和观察到的值。RMSE表示预测值与观测值之间差异的第二样本矩的平方根或这些差异的二次均值。当对用于估计的数据样本执行计算时，这些偏差称为残差，在样本外进行计算时称为误差（或预测误差）。RMSE用于将不同时间的预测误差的大小聚合成单个预测功率测量。RMSD是精确度的度量，用于比较特定数据集的不同模型的预测误差，而不是数据集之间的预测误差，因为它与比例相关。

RMSE始终是非负的，值0（实际上从未实现）表明数据非常合适。通常，较低的RMSE优于较高的RMSE。

通过对各个degree的RMSE的比较，可以发现：该组数据在多项式回归时，取次数为6或者9时，模型建立最佳。

当degree=6时：

女生数据：0.79

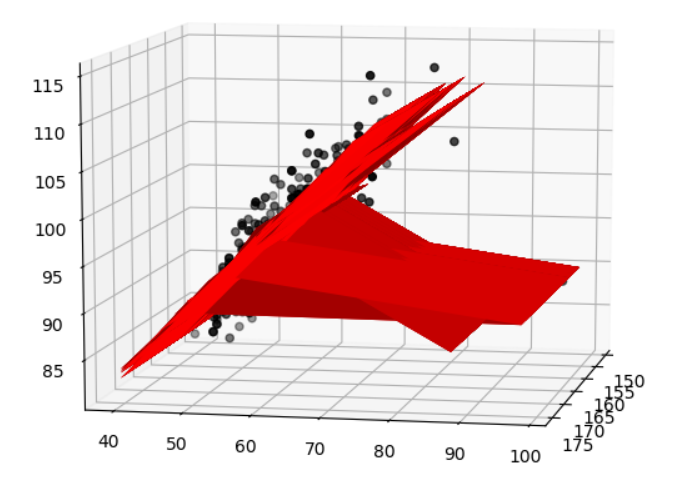


图52

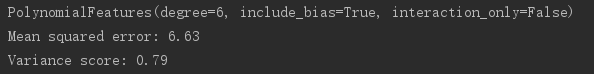


图53

男生数据：0.80

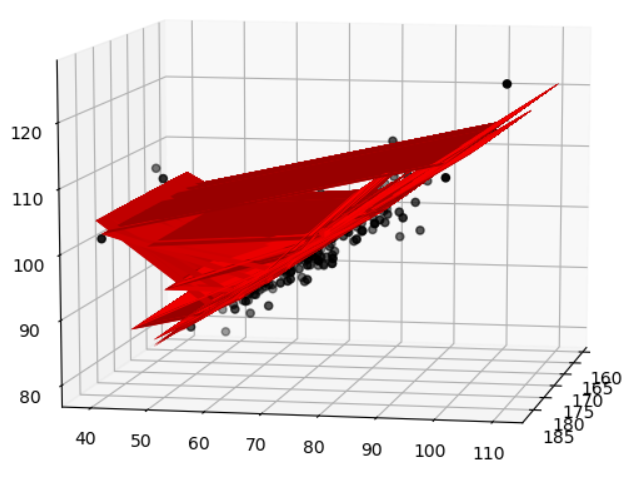


图54

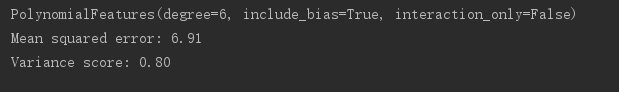


图55

当degree=9时：

女生数据：0.79

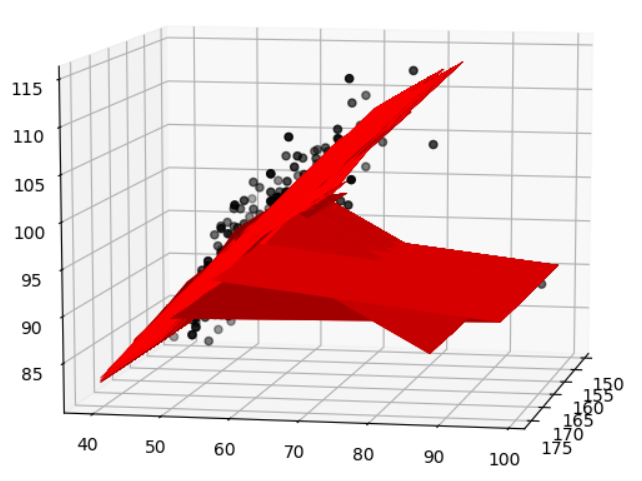


图56

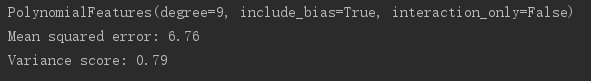


图57

男生数据：0.81

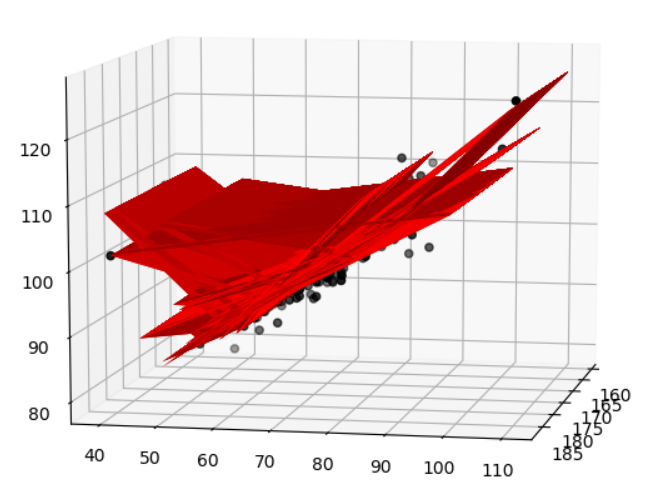


图58

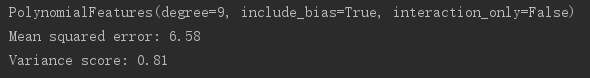


图59

#### 总结

当已知[身高,体重]，预测[紧臀围]时，本次实验结果所呈现的回归方法评分都比较相近。可能原因有以下几个：

1. 在数据清洗时，抛弃了太多异常值，导致数据离散程度较小，如上图所显示的那样，呈现出比较密集的散点，这样的分布可能会导致回归算法不能有效的得出最优的曲线。
2. 本次数据集虽然总数居高达5998条，但是经过数据清洗之后，剩下了仅仅3000多条。不充分的数据，也会导致算法得到的结果有相异度，但是相异度太小，无法体现。当数据集的数据记录足够多，展现出足够丰富的属性值时，各个回归方法的优点才会各自体现。
3. 本次数据集的记录是反映人体的量体数据，众所周知，人体的量体数据比较集中，会导致无法有效拟合等问题。

与此同时，我们也可以发现，LARS Lasso回归算法并不适合于低维度的数据分析，当需要处理多维数据集时，这个算法可能才会体现出优势。而岭回归算法的使用准确率远远高于其他算法。

### 问题四：预测相关数据的准确性

本次项目所使用的所有回归算法，都带有predict函数，可以通过train集得出对应最优回归曲线，然后使用test集对得出的曲线进行验证，从而得到预测值和正确率。

在上文数据清洗时，已经运用到predict函数，来对数据是否具有不可猜测的错误性进行验证。当然，在绝大多数其他实际案例中都可以通过predict函数来对未知数据的属性进行预测。本项目将通过员工（已知性别）的身高、体重，预测未知的紧胸围和紧臀围。

当将test数据带入算法时，会得到对应的一个predict集，在matplotlib包中调用plot和scatter方法，画出x=y和散点。散点坐标为（预测值，测试值），当散点越靠近直线则表示预测值越精准。下面为各算法对于不同的随机数据集展现出的预测情况。

#### [Ordinary Least Squares](https://en.wikipedia.org/wiki/Ordinary_least_squares)普通最小二乘法

女生预测结果：

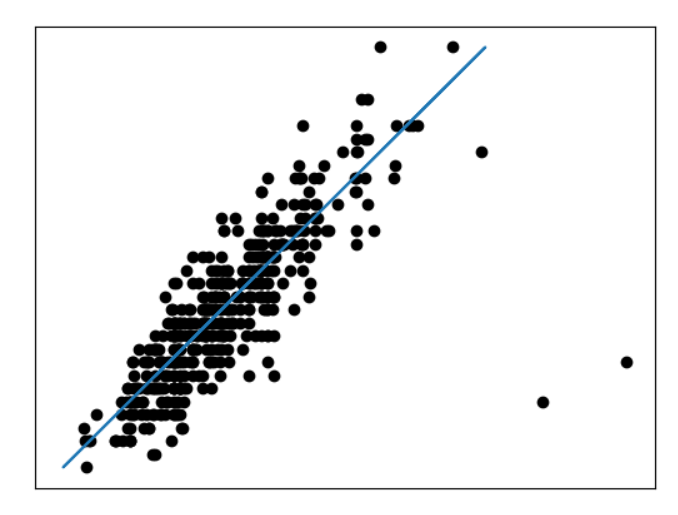


图60

男生预测结果：

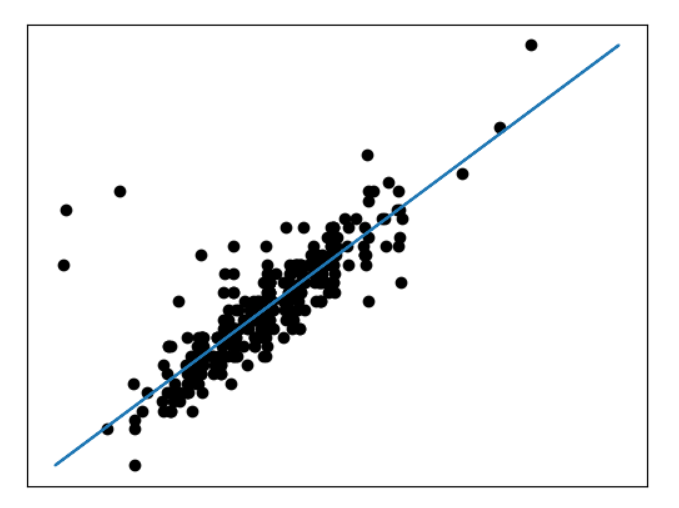


图61

使用cross\_val\_predict函数，可以简易地对数据进行交叉验证，并获得平均预测值。cross\_val\_score能够提供各次交叉验证后的均方差评分。

这里使用参数cv=10，进行十折交叉验证，将数据分为10份，每次分别将1份作为测试集，其余作为训练集进行验证。利用平均预测值与全体数据计算得到十折交叉验证的综合均方差评分。

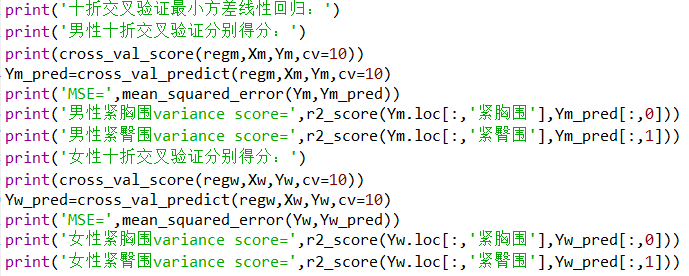


图62

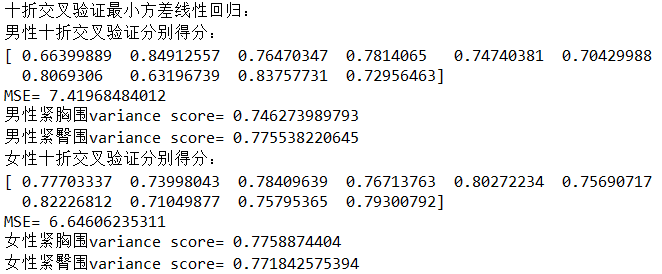


图63

十折交叉验证下，男性、女性的紧胸围、紧臀围平均预测值与实际值如图64：

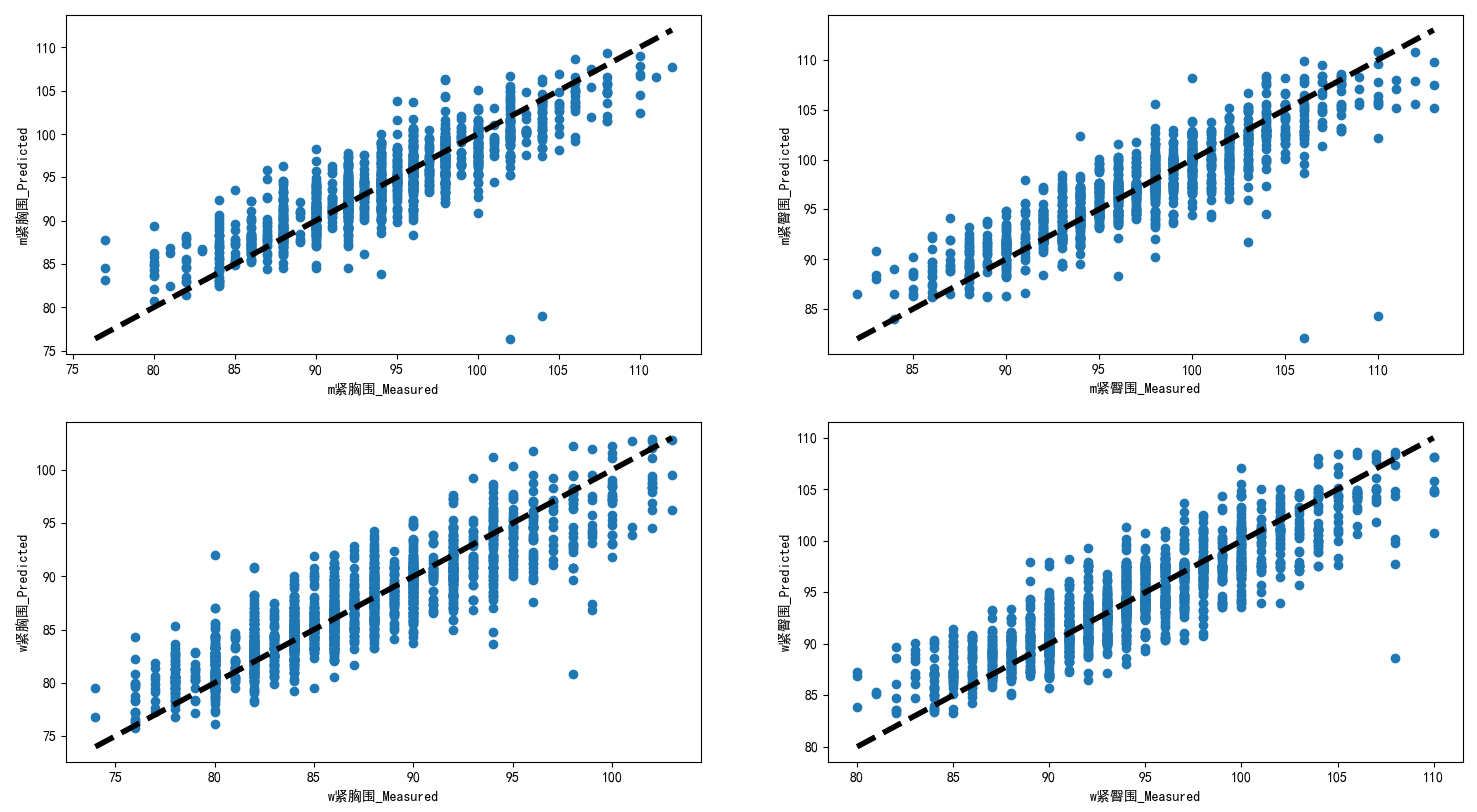


图64

#### [Ridge Regression](https://en.wikipedia.org/wiki/Ordinary_least_squares)岭回归（吉洪诺夫正规化）

女生预测结果：

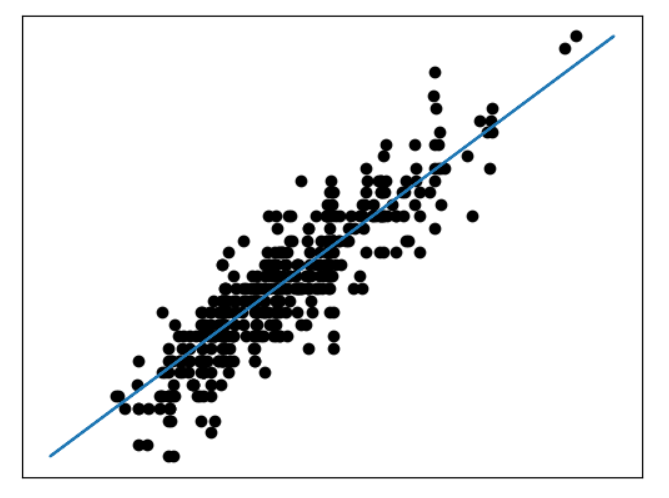


图65

男生预测结果：

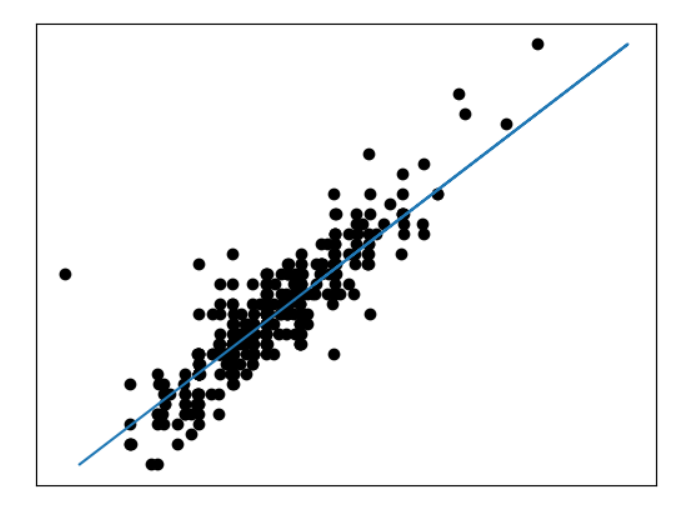


图66

对岭回归方法进行十折交叉验证如图67：

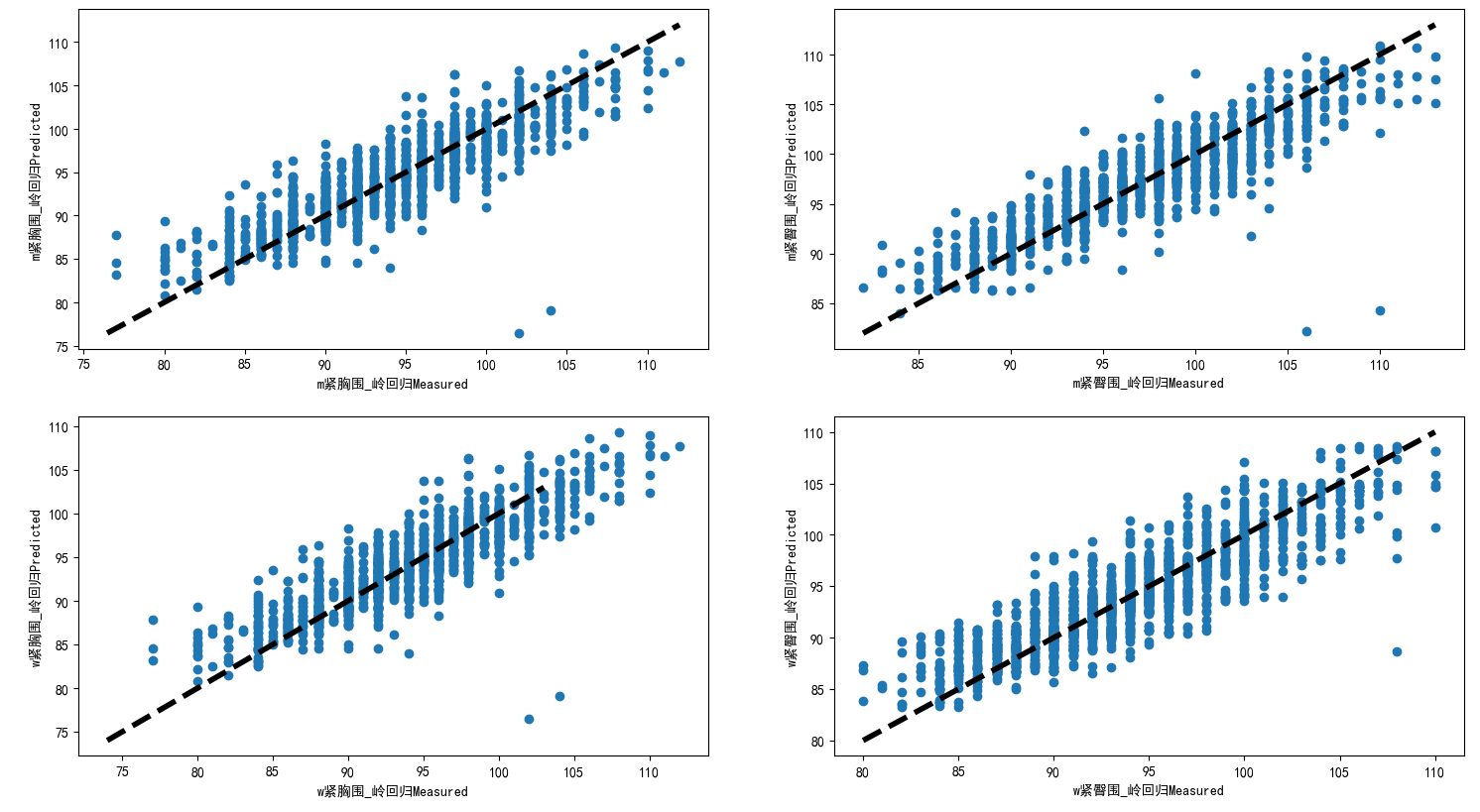


图67

#### [Lasso](https://en.wikipedia.org/wiki/Lasso_(statistics))套索回归

女生预测结果：

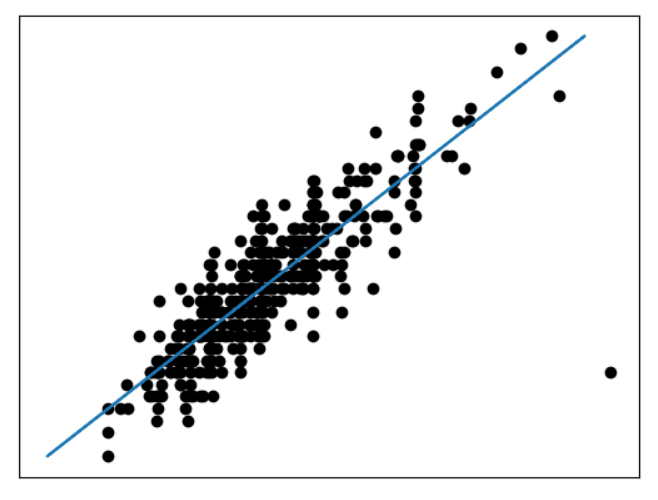


图68

男生预测结果：

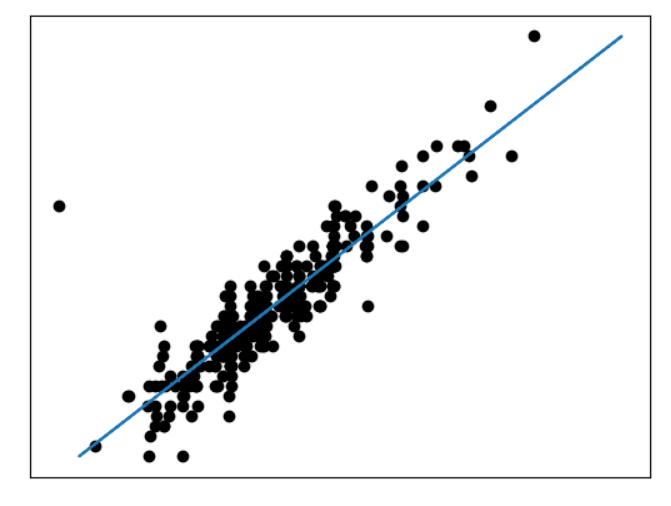


图69

#### [Bayesian Ridge Regression](https://en.wikipedia.org/wiki/Ordinary_least_squares)贝叶斯岭回归

女生预测结果：

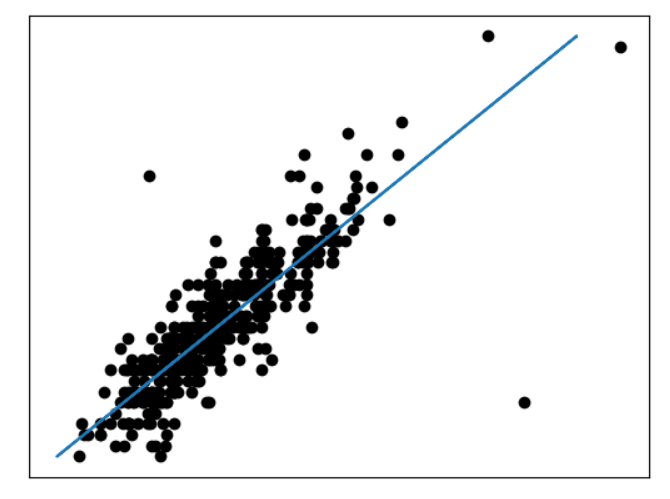


图70

男生预测结果：

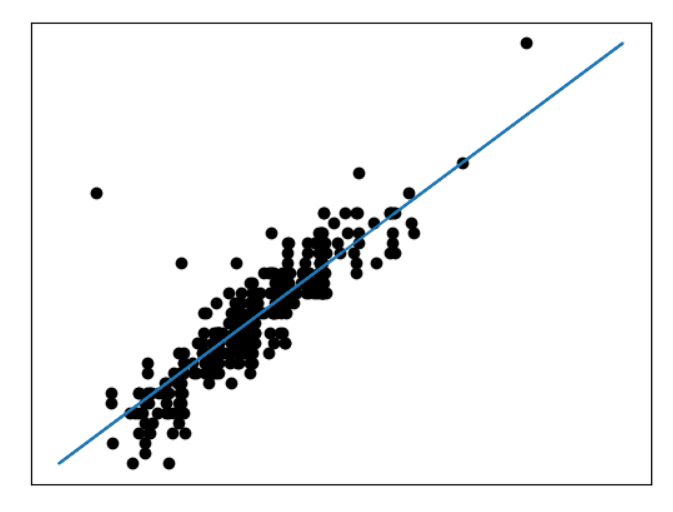


图71

#### [LARS](https://en.wikipedia.org/wiki/Least-angle_regression) 最小角Lasso回归

女生预测结果：

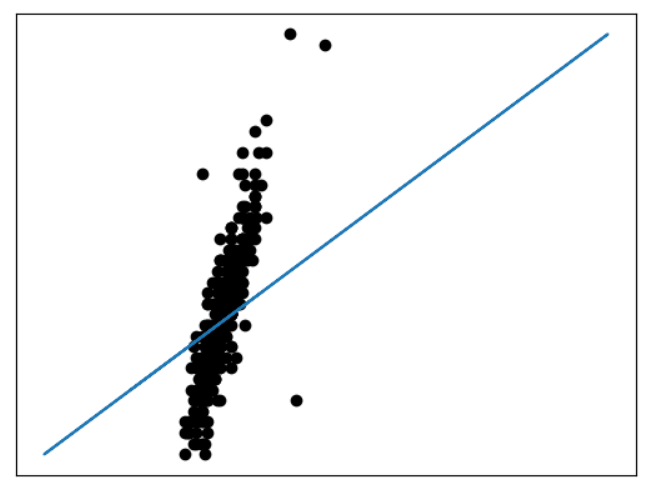


图72

男生预测结果：

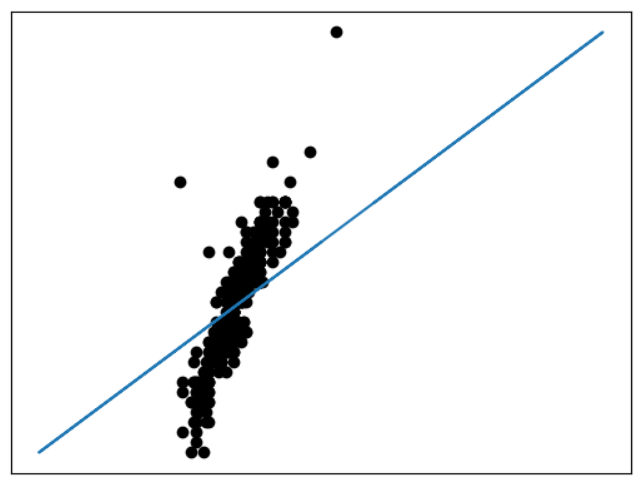


图73

#### [Polynomial regression](https://en.wikipedia.org/wiki/Polynomial_regression)多项式回归（degree为9时）

女生预测结果：

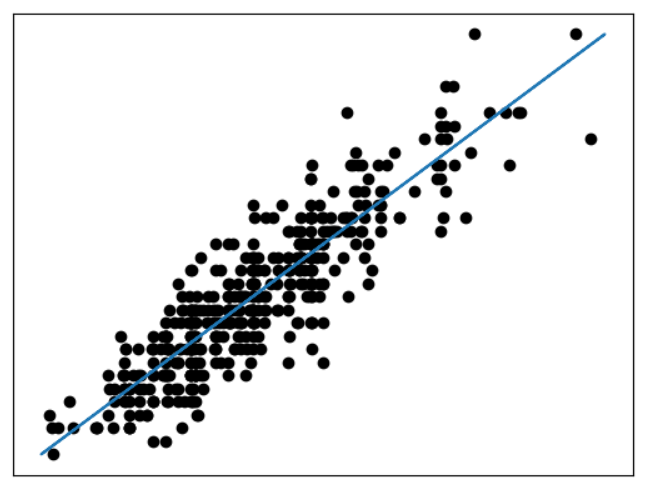


图74

男生预测结果：

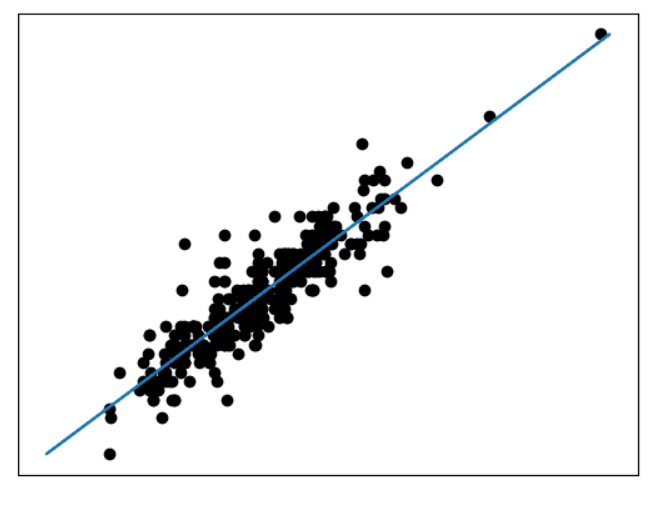


图75

#### 总结

由上述图片，可以更加肯定的得出这样的结果：回归算法的预测准确率和算法评分是呈正相关的关系。

## 数据挖掘的应用前景和展望

目前数据挖掘得到了充分而迅速的发展，现在的数据挖掘热点在于各种大数据的精确挖掘，例如网站上的数据挖掘，如上文提到的电子商务内容；对生物的基因信息进行数据发掘；从大批量文本中进行数据挖掘等。未来的发展趋势很大程度上也会沿着这三个方向进行进一步的发展和完善。我国的数据挖掘尚不成熟，处于发展阶段，但各行各业对于数据挖掘的重视程度正逐步提高，对数据挖掘的研究和应用也越来越广泛。

### 总结

数据挖掘是一个常见而好用的数据处理工具，在如今的电子商务时代和大数据时代并行的时代背景下显得尤为重要，目前国内外对数据挖掘的热度正逐渐上升，多个数据量大且对数据处理需求较大行业已经开始应用数据挖掘来进行辅助决策，数据处理在日常的数据分析中也开始逐步应用。但是就目前而言数据挖掘还存在着许多缺憾和不足，例如在数据量非常大时，数据发掘的速度和效率亟待提高、动态数据的挖掘方式不足、对多数据类型的挖掘方法不够等。作为数据的分析者，必须理解和掌握数据挖掘的原理和使用方法，熟练和合理的应用数据挖掘，从而能解释和应用最终的到的数据挖掘结果，使得数据挖掘能真正有利于自己的工作，服务于行业。