

探究心肺功能与体测长跑成绩之间的关系

陈思远 16307110003

一、实验背景及意义

学校运动会在 5 月月底举行，师生们的运动热情十分高涨。人在激烈运动时，身体各项机能都会在短时间急剧提升，还有体液中激素水平的上调。这些身体机能的变化都可以使人体在短时间内获得比平常更强的运动能力。而其中的心肺功能的优劣更是对中长跑运动成绩的提高有着直接的关系^{[1][2]}。目前，国内该领域研究对心肺功能指标的运动成绩预测能力研究较少，更多的研究集中于如何合理地评估运动员间的心肺功能差异^{[3][4]}。本实验小组对剧烈运动前后的心脏收缩压、舒张压、平均心率、潮气量、呼吸频率等心肺功能指标进行检测，结合实验对象的年龄身高体重等基本参数得出年轻男性群体长跑成绩与检测指标相关性。本实验将使用 R 语言编程，建立预测模型，并通过测得的心肺功能相关数据对预测模型进行训练，从而获得能够较科学地预测学生长跑成绩的模式。本组希望能通过一些易于获得的心肺指标对学生长跑成绩建立一个较好的估计，使同学们通过这些易于获得的指标对自己的中长跑能力有一个大致了解。

二、实验原理

1.生理学原理

人类的身体机能与其运动能力直接挂钩。先前研究表明，心肺功能的优劣与中长跑运动成绩有显著的相关性。一般来说，心肺功能越强，其中长跑能力越强，表现出的中长跑成绩更加优异^{[1][2]}。

常用的运动生理学心肺功能测试指标如下：

心功能：射血分数（EF），心排血量（CO），心搏指数（CI），心搏出量（SV），峰充盈率（PFR），峰充盈时间（TPFR），快速充盈率（FFR）。

肺功能：肺活量（VC），肺总量（TCL），残气量（RV）/肺总量（TCL），最大通气量（MVV），一秒用力呼气量（FEV1）。

通气功能检查：肺通气量（又分为每分钟静息通气量 VE 和最大通气量 MVV），用力肺活量（FVC），最大呼气中段流量（MMEF，MMF），肺泡通气量（VA）。

换气功能检查气体分布；通气/血流比值；弥散功能。

小气道功能检查：闭合容积（CV）；最大呼气流量-容积曲线（MEFV,V-V 曲线）；频率依赖性肺顺应性（FDC）。

在实验中，由于受制于实验条件，本组选择了复旦大学男性学生作为研究对象。对他们静息状态下心率、收缩压、舒张压、呼吸频率、潮气量、一秒用力吹起量、肺活量及剧烈运动后心率、收缩压、舒张压、呼吸频率、潮气量等指标进行检测，并向被试获得年龄、身高、体重、1000m 成绩和平均每周运动时间等信息，分析其与 1000 米跑成绩的相关性，并以此数据对预测模型进行训练，进而获得较科学地预测模型。

2.数据获取

2019 年 5 月 6 日至 5 月 22 日，以校内男性学生为实验对象进行 50 余组试验。统一令被试以每分钟 80 拍的频率进行抬腿跳运动，持续时长一分钟。抬腿跳运动使用实验室门口木板长椅限制被试的抬腿高度。

利用血压仪、肺活量检测仪检测学生的舒张压、收缩压、心率与肺活量、一秒吹气量；在抬腿跳运动前后利用弹力收缩检测器和静脉压检测器结合具有模电转换装置的计算机和

LabChart 软件绘制学生的肺部收缩周期图与心电图,并计算出呼吸频率、潮气量与平均心率。

3.模型构建

通过压缩估计的方法建立线性模型。压缩估计该方法基于全部 p 个预测变量进行模型的拟合。该方法可以将估计系数往零的方向进行压缩,通过系数缩减减少方差。通过不同的系数缩减方法,某些回归系数可以被缩减为零,也由此进行变量的选择。

岭回归,即通过最小化下式来确定每个变量前的系数:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^p \beta_j^2.$$

当预测变量中的大多数与相应变量有较好的相关性时,岭回归的方法能有较好的预测效果。但当多数预测变量与响应变量相关性较差时,岭回归的拟合效果较差,且变量过多可能对模型解释造成一些困难。

Lasso,通过最小化下式来确定每个变量前的系数:

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^p |\beta_j|$$

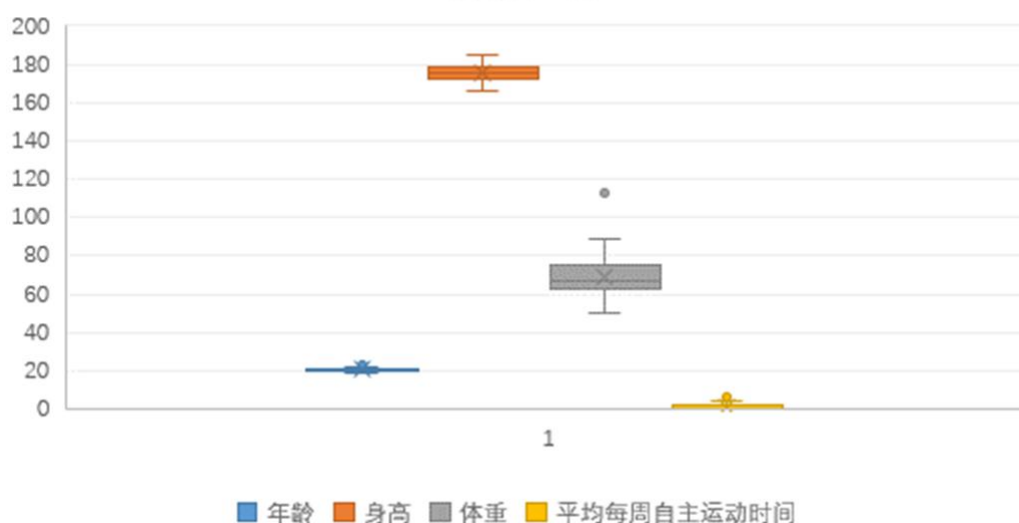
Lasso 方法可以将预测变量中相关性较差的直接压缩为 0,在多数预测变量与相应变量相关性较差时有更好的拟合效果,且排出了相关性差的变量使模型的解释更为简单。但当多数预测变量与相应变量高度相关时,其效果不如岭回归。

若结果明显不符合线性假设,则考虑使用其他非线性统计方法,如多项式回归。

三、实验对象

复旦大学本科生 55 名,其中女生 1 名、数据不全 3 名,排除后有效数据共 51 份,平均年龄 20.61 岁,平均身高 175.24cm,平均体重 68.49kg,平均每周运动时间约 1.04 小时。

基础数据箱形图



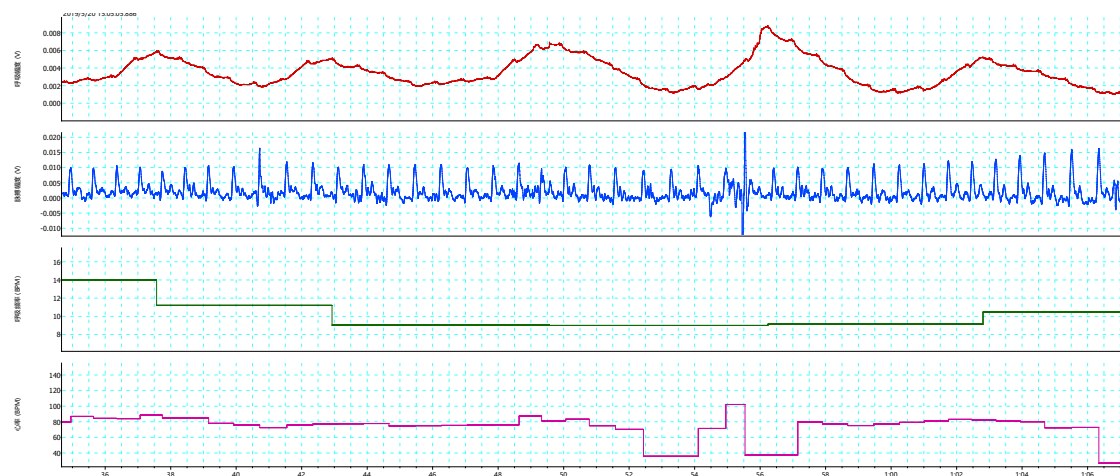
四、实验步骤

- 1、受试者在静息状态下用血压计于左臂测量血压，记录心率、收缩压和舒张压。
- 2、受试者在静息状态下测肺活量并记录，然后在按下秒表的同时用力吹气，1秒后停止，记录1秒用力呼气量。
- 3、设置 labchart 参数，将呼吸换能器的胸带绑于胸腔偏上处（若无法获得呼吸曲线或呼吸曲线不明显，可将换能器下移至横膈膜处），同时将脉搏换能器绑于右手中指根部（防止运动时脱落），静坐，手放于桌上，记录静息时呼吸曲线与脉搏曲线。
- 4、保持换能器位置，运动1分钟（以80BPM的速度利用长凳做抬腿运动），然后立即用血压计测量运动后血压，记录运动后心率、收缩压、舒张压、呼吸曲线与脉搏曲线。
- 5、利用呼吸曲线计算运动前后受试者呼吸频率、潮气量（用呼吸曲线的平均幅值代替）。
- 6、登记受试者姓名、年龄、身高、体重、1000米成绩和平均每周自主运动时间等基本信息。
- 7、后续数据处理与分析。

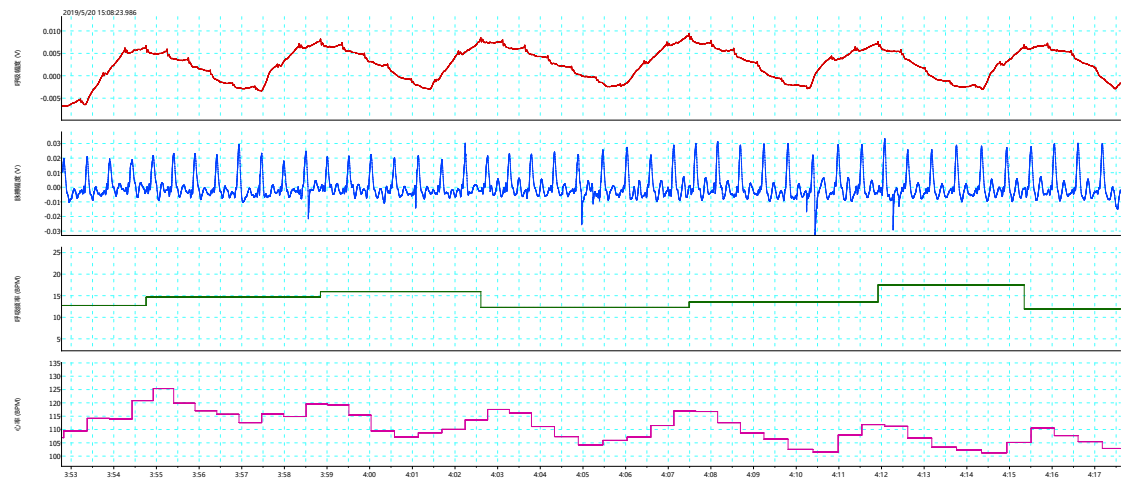
五、实验结果

1.运动前后心肺生理变化

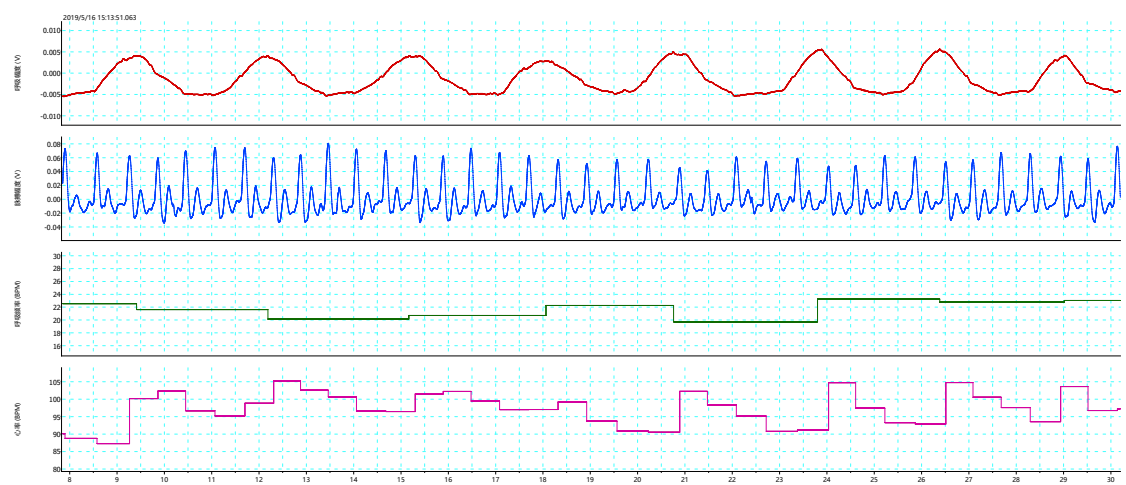
根据被试的1000m成绩，将被试大致按照成绩分为1000m用时 $<4\text{min}$ 和 $\geq 4\text{min}$ 两组。观察各组组内运动前后图像，对比发现：运动前呼吸图线较为平缓、峰值较低，峰间距较远；运动后呼吸图线紧促、峰值较高，两峰间距较近。运动前脉搏图线规律缓和，运动后脉搏较为急促杂乱。这表明每位被试者运动前后呼吸与心跳活动均有明显增强，这一变化的强度一定程度上反映了被试者的心肺功能情况。



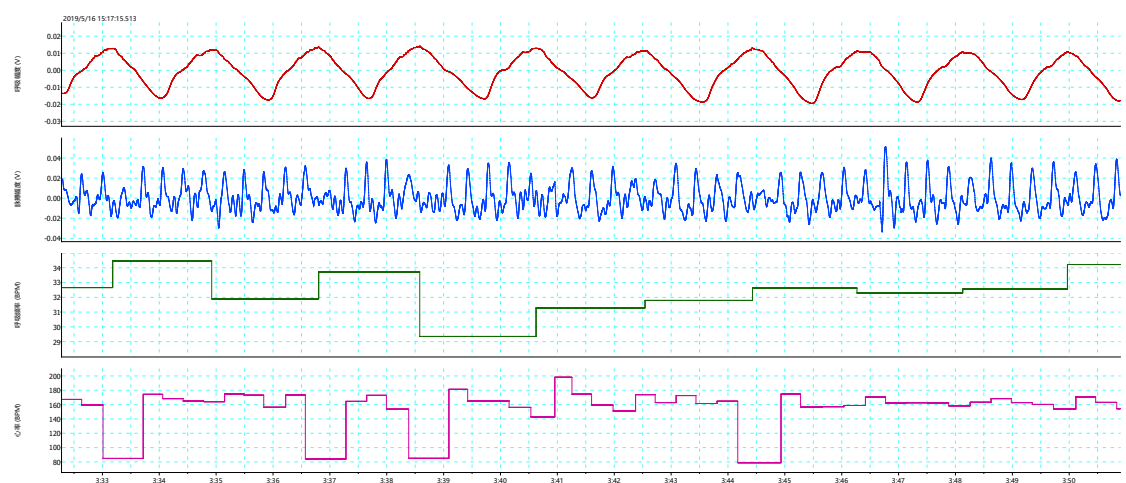
(a)



(b)



(c)



(d)

图 1 心肺功能曲线

每张图片从上到下分别为呼吸幅度、脉搏强度、呼吸频率、脉搏频率。(a) 与 (b) 为 1000m 用时小于 4min 的某被试运动前与运动后的数据。(c) 与 (d) 为 1000m 用时大于 4min 的某被试运动前与运动后的数据。

下表对两个组别被试的生理指标作一个初步的比较：1000m 成绩<4min，运动能力更强的小组静息心率明显更低，在统计学上有显著差异，且该组运动前后潮气量的变化较1000m 成绩≥4min 的被试更大，表明该组学生在运动后增大呼吸幅度的能力更强。

表 1 各组学生测试数据

组别	静息状态（均值）			运动后（均值）		
	潮气量 (V)	呼吸频率(BPM)	心率 (BPM)	潮气量 (V)	呼吸频率 (BPM)	心率 (BPM)
<4:00	0.0097	15.3	70	0.0369	21.7	124.2
>=4:00	0.0105	16.5	82.2***	0.0200*	22.9	128.2

对每组数据在上下两个组别之间使用单因素方差检验。*表示 P<0.05,**表示 P<0.01,***表示 P<0.001。

2.数据处理与预测分析

1)共线性检验

计算目标变量的方差膨胀因子（vif）

年龄	身高	体重	BMI	平均每周运动时间
1.93	47.00	304.50	303.01	2.81
静息心率	静息收缩压	静息舒张压	静息呼吸频率	肺活量
2.16	4.00	5.67	3.21	42.20
肺活量/体重	1秒用力呼气量	运动后心率	运动后收缩压	运动后舒张压
48.04	1.41	2.66	2.48	3.49
运动后呼吸频率	运动前后潮气量差			
3.48	1.33			

经验判断方法表明：当 $0 < VIF < 10$ ，不存在多重共线性；当 $10 \leq VIF < 100$ ，存在较强的多重共线性；当 $VIF \geq 100$ ，存在严重多重共线性。故可以判断，在数据中存在共线性的情况，在之后的处理中不能直接进行简单线性模型的拟合，考虑使用两种压缩估计的方法。

2)数据标准化

由于预测变量的尺度会对压缩估计产生影响，所以需要先对数据进行标准化，具体变准化过程如下式所示。

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

3)岭回归

使用岭回归的方法，将所有变量前的系数部位零的压缩，系数下降越快的变量与学生一千米成绩之间的相关性越差（图 2A）。使用 3-fold 的交叉验证选择压缩系数 λ ，当 $\lambda = 17.2$ 时，该预测模型有最小的 MSE，约为 $304.7s^2$ ，即平均的测试误差约 $\pm 17.5s$ （图 2B）。我们可以明显地从图 A 中看到，许多变量的系数都被快速压缩，在相关系数较小的区域有大量系数存在，在这种情况下岭回归不是一个较好的估计方法。

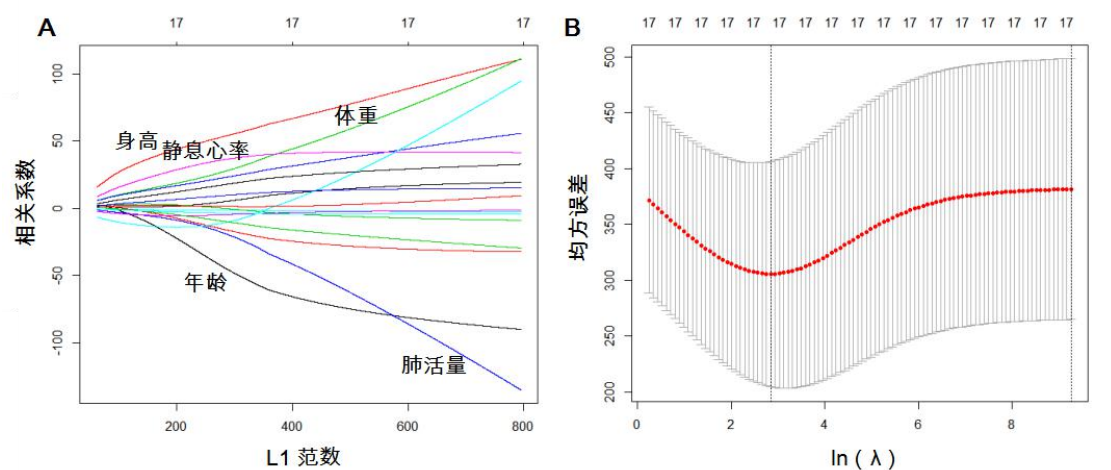


图2 岭回归

(a) 各变量相关系数的压缩情况。(b) 使用 3-fold 交叉验证挑选合适的 λ ，并得到使用该 λ 时模型的测试均方误差。(即从数据集中随机抽取 2/3 进行模型建立，1/3 用作测试)

岭回归得到的线性模型各变量前的系数如下表所示：

变量	系数
截距	172.0567389
年龄	-14.31594923
身高	39.39920243
体重	15.80691981
BMI	14.64664484
平均每周运动时间	-1.021912898
静息心率	24.93480663
静息收缩压	1.279342697
静息舒张压	-4.361740024
静息呼吸频率	3.065945416
肺活量	-6.510349341
肺活量/体重	-13.58758479
1 秒用力呼气量	-5.308229436
运动后心率	10.30768371
运动后收缩压	2.11454678
运动后舒张压	-3.10400142
运动后呼吸频率	5.657605664
运动前后潮气量差	-2.178704556

从岭回归的结果中我们大体可以看出，身高、体重、BMI、静息心率、运动后心率、运动后呼吸频率越大，学生一千米用时更长；年龄、肺活量、肺活量/体重、1 秒用力呼气量越大，学生一千米用时更短。但是，由于岭回归的结果中有过多的冗余变量，对模型的解释产生不便。

4) Lasso 回归

Lasso 方法与岭回归类似，也是一种压缩估计方法，但是其能将相关性较差的预测变量前的系数压缩为 0，排出这些系数的影响。在大量预测变量与响应变量相关性不好时，Lasso 方法能有更好的估计效果，在我们的数据中可能有更好的应用。随着各变量被不断压缩，静息心率、体重、运动后呼吸频率、肺活量/体重四项是最后被压缩为 0 的四个变量，前三者的系数为正，最后一个的系数是负数（图 3A）。这也就意味着，静息心率越高、体重越大、运动后呼吸频率越高的同学一千米用时会更长，而肺活量与体重的比值越大，同学的一千米用时会更短。这与人们的日常经验相符。同样地，使用 3-fold 交叉验证进行压缩系数 λ 进行筛选。当 $\lambda = 2.2$ 时，lasso 方法的预测模型有最小的 MSE，约为 $273.1s^2$ （图 3B），比起之前的岭回归有一定程度的下降。平均测试误差约为 $\pm 16.5s$ 。

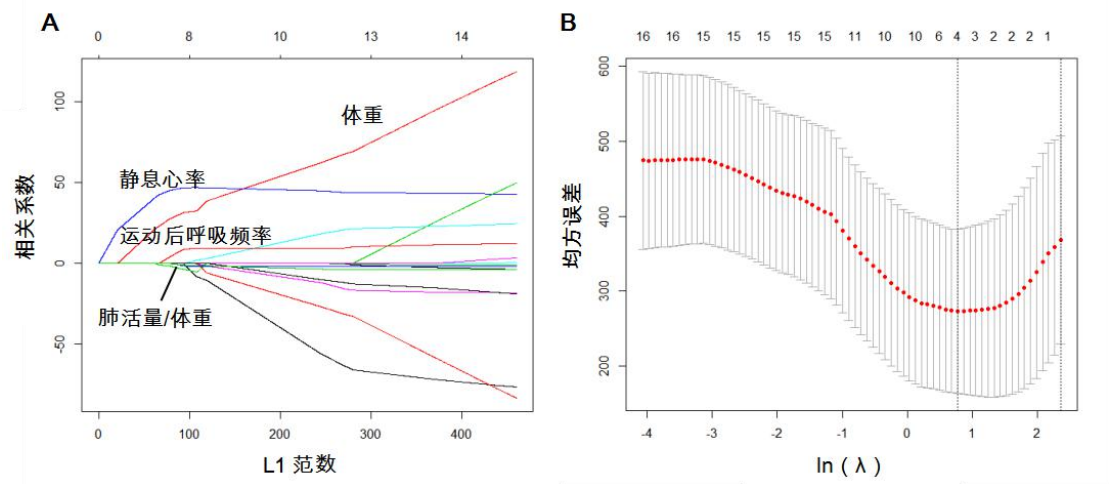


图 3 Lasso 回归

(a) 各变量相关系数的压缩情况，其中最后被压缩为 0 的四项变量为静息心率、体重、运动后呼吸频率、肺活量/体重。(b) 使用 3-fold 交叉验证挑选合适的 λ ，并得到使用该 λ 时模型的测试均方误差。（即从数据集中随机抽取 2/3 进行模型建立，1/3 用作测试）

Lasso 得到的线性模型各变量前的系数如下表所示：

变量	系数
截距	166.0830696
年龄	0
身高	0
体重	27.34800015
BMI	0
平均每周运动时间	0
静息心率	44.93500474
静息收缩压	0
静息舒张压	0
静息呼吸频率	0
肺活量	0
肺活量/体重	-1.457063001
1 秒用力呼气量	0

运动后心率	0
运动后收缩压	0
运动后舒张压	0
运动后呼吸频率	3.728413562
运动前后潮气量差	0

5) 非线性模型

然而，lasso 方法虽然比起岭回归有所改进，但其预测误差仍在一个较高的水平。我们猜想这可能是由于这些变量与一千米成绩之间的关系是非线性的。线性假设可以用残差与相应变量间作图来检验，理论上一个符合线性假设的模型，其残差与相应变量之间无关，数据点在图中应随机分布。然而，如图所示，残差与相应变量之间存在着一定的联系，近乎线性，当响应变量越大，残差值越小，即说明预测变量与响应变量间的线性关系较弱（图 4）。

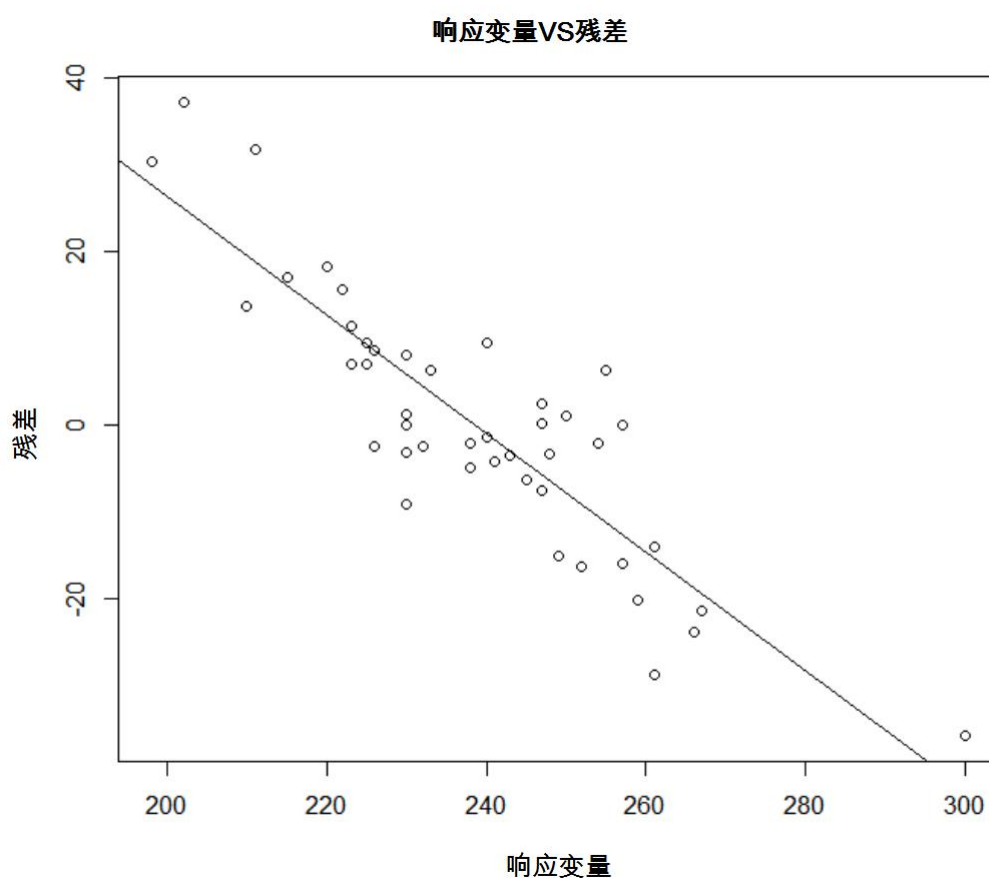


图 4 响应变量与残差作图

残差值与相应变量的关系如图所示，图中直线为数据点的线性拟合结果。

因此，考虑对相应变量进行非线性的拟合。之前 lasso 方法已经为我们挑选出了与响应变量相关性较好的 4 组预测变量，分别为体重、静息心率、肺活量/体重及运动后呼吸频率，我们尝试对这四个变量进行多项式回归，且同样使用 3-fold 的交叉验证方法生成测试集（即从数据集中随机抽取 2/3 进行模型建立，1/3 用作测试），对拟合的效果进行评估。这 4 组变量之间几乎没有线性关系。

当使用二次多项式进行预测时，MSE 为 $141.1s^2$ 平均测试误差为 $\pm 11.9s$ ，较之前有显著

下降。模型的 R^2 为 0.65，经过 F 检验得到 p 值为 0.0018。该二次多项式拟合的预测模型与预测变量间存在显著关联。拟合得到的函数式具体如下：

学生一千米成绩 = $406.735003 - 3.675683 * \text{体重} + 0.025111 * \text{体重}^2 + 1.870907 * \text{静息心率} - 0.006906 * \text{静息心率}^2 - 2.908465 * \text{肺活量/体重} + 0.018200 * (\text{肺活量/体重})^2 - 2.482062 * \text{运动后呼吸频率} + 0.059239 * \text{运动后呼吸频率}^2$ 。（数据未经过标准化，使用单位为 kg, BPM, ml/kg 的原始数据即可）

然而，当使用三次多项式预测时，MSE 不减反增，为 $247.1s^2$ ，平均测试误差为 $\pm 15.7s$ 。这可能是由于随着次数增加，出现了过拟合的现象。

六、讨论与总结

本组通过实验，得到了 4 个变量对男生 1000m 成绩有较显著影响，分别为体重、静息心率、运动后呼吸频率、肺活量/体重。其中，体重、静息心率、运动后呼吸频率越大，1000m 用时更长；肺活量/体重越大，1000m 用时更短。此外，通过二次多项式拟合，获得了平均测试误差为 $\pm 11.9s$ 的预测模型。二次多项式得到的模型也可以较好地解释我们的数据。如体重项， $-3.675683 * \text{体重} + 0.025111 * \text{体重}^2$ ，在学生体重接近 73kg 时取到最小值，过于瘦弱或肥胖的学生跑 1000m 的用时都会更长。我们的结果可以使学生在大三、大四两年没有体育锻炼的情况下通过简单的心肺功能指标对自己现阶段的运动能力有大致的判断，有利于防止身体素质快速下降的现象出现。

然而，我们的实验仍存在一些不足之处。首先，在中长跑运动员中对心肺功能的研究多是在体育学院或更为专业的运动员中完成^{[3][4]}，学生跑 1000m 时的成绩可能会受到如跑步姿势等更多的因素影响，且我们无法判断该成绩是否是某位同学尽全力的 1000m 成绩，这会直接影响到我们的实验。其次，虽然我们同时记录了学生运动时的心肺数据，但其随着运动有明显波动，对心率变化的潜伏期等指标判断带来困难。最后，作为一个统计学的实验，可能还需要进一步增加样本量以获得更好的预测结果。

参考文献：

- [1] Lu, K. D., Bar - Yoseph, R. , Radom - Aizik, S. and Cooper, D. M. (2019), A New Approach to Estimate Aerobic Fitness Using the NHANES Dataset. Scand J Med Sci Sports.
- [2] Marlus Karsten, Gustavo S. Ribeiro, Mateus S. Esquivel, Darlan L. Matte(2018), The effects of inspiratory muscle training with linear workload devices on the sports performance and cardiopulmonary function of athletes: A systematic review and meta-analysis. Physical Therapy in Sport.
- [3] 唐汝安. 对少年中长跑运动员心肺功能的判别分析[J]. 辽宁体育科技, 1981(Z1):35-38.
- [4] 李艳斌, 贾巍. 少年中长跑运动员心肺功能特点的研究[J]. 沈阳体育学院学报, 2010, 29(5): 64-67.