新冠疫情等因素对学生学习效果的影响

黄河源 201800820087 数据科学与人工智能实验班

1. 引言

新冠疫情于 2019 年底爆发，2020 年初各行业开始尝试远程办公与教学。线上办公、学习等新方式减少了人与人之间的接触，可以有效降低疫情传播速率和感染率，保障公民的生命安全。但与此同时，线上办公学习也会带来“人员无法管控”从而导致办公效率降低、学习质量差等难题。

为了研究线上线下教学模式等因素对学生学习效果的影响，美国波特兰州立大学研究生助理 Dylan Bollard 根据俄勒冈州当地的人口统计学数据，模拟了新冠前后共 6 个学期的真实情况，生成了当地的面板数据集。

根据本文模型的分析结果，由于网络教学，校内学生的学习表现大大下降，且呈持续降低的趋势。当疫情严重时，教育机构应加强对线上教学的管控，如利用点名保障学生的出勤率，及时收取并批改作业、增加定期测试环节以保证学生的课业进度。当疫情被有效抑制后，教育机构应及时恢复线下教学，保证教学质量，并保持研究如何提高线上教学的监督效果，以应对疫情的反复。

2. 理论分析与研究思路

学生学习效果的影响因素属于社会学范畴，包括学校的师资条件、学生的家庭条件、性别、种族、父母学历、学生性格等，而新冠疫情的发生导致“教学方式”变成了备受关注的影响因素。

为了研究面对面教学和线上教学的质量差异，分析学习效果的影响因素，可将学生的期末考试成绩作为被解释变量，反映教学效果的好坏，各种影响因素作为解释变量，建立回归方程。通过分析回归方程拟合效果是否显著，我们可以研究各因素对学习成绩的正负向影响，提出合理的优化方案。

3. 数据及处理

本论文使用的数据集为美国波特兰州立大学提供的开源面板数据集[1]，旨在模拟俄勒冈州地区 6 个学期的真实数据，样本数目为 8400，为大样本数据集。前三个学期为新冠疫情封锁前，各学校采用线下面对面教学模式。后三个学期为新冠疫情封锁后，各学校采用线上教学模式。

数据集中各变量如下：

	变量类型	变量解释
学号	整数	学生 ID，用作数据索引
学校条件	虚拟变量	0 表示富裕学校，1 表示贫穷学校

年级	整数	学生所在年级（6-12）
性别	虚拟变量	0 为女性，1 为男性
是否确诊新冠肺炎	虚拟变量	0 为没有患病，1 为确诊
是否申请学校免费午餐	虚拟变量	0 为未曾申请, 1 为申请了免费或降价午餐
家庭电脑数目	整数	学生家里有几台电脑
第几学期	整数	[0,1,2]表示线下教学，[3,4,5]表示线上教学
家庭成员数	整数	包括父母、兄弟姐妹在内的家庭总人数
用在儿童身上的家庭收入	整数	家庭收入中用在孩子身上的金额数目
父亲学历	整数	0 表示高中肄业, 1 表示高中学历，2 表示本科学位，3 表示硕士学位，4 表示博士学位
母亲学历	整数	0 表示高中肄业, 1 表示高中学历，2 表示本科学位，3 表示硕士学位，4 表示博士学位
阅读考试分数	浮点数	学生在 reading 阅读考试中得到的期末成绩
写作考试分数	浮点数	学生在 writing 写作考试中得到的期末成绩
数学考试分数	浮点数	学生在 math 数学考试中得到的期末成绩
俄勒冈州阅读考试分数	浮点数	学生在俄勒冈州 reading 阅读考试中得到的期末成绩
俄勒冈州写作考试分数	浮点数	学生在俄勒冈州 writing 写作考试中得到的期末成绩
俄勒冈州数学考试分数	浮点数	学生在俄勒冈州 math 数学考试中得到的期末成绩

美国举行州考试和学校考试，旨在不考虑学校之间考试出题差异的影响，类似全国统一高考卷，研究整个州的学习效果。

在数据处理方面，用在孩子身上的家庭收入平均值为 84290.2225 美金，对其进行10^{-4}尺度放缩，以万为单位，使估计出的系数可读性提高。用阅读、写作、数学考试的分数之和作为被解释变量，突出学习的总效果，减少因偏科等因素带来的影响。

4. 计量经济模型与估计方法

设被解释变量Y_1为学生学校阅读、写作、数学考试的分数之和，解释变量

$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}$分别为学校条件、年级、性别、是否确诊新冠肺炎、用于儿童身上的家庭收入、是否申请学校免费午餐、家庭电脑数目、家庭成员数、父亲学历、母亲学历、学期。

建立以下多元线性回归模型：

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_{11} X_{11t} + u_i$$

使用 OLS 进行回归参数的估计，首先需要进行基本假设：

对于任意给定的$i = 1, \ldots, n$ 有以下基本假定成立：

$$E(u_i) = 0 \text{ 零均值假定}$$
$$Var(u_i) = \sigma^2 \text{ 同方差假定}$$
$$Cov(u_i, u_j) = 0 \text{ 无自相关假定}$$
$$Cov(u_i, X_{ji}) = 0 \text{ 随机扰动项与解释变量不相关假定}$$
$$Rank(X) = k \text{ 无多重共线性假定}$$
$$u_i \sim N(0, \sigma^2) \text{ 正态性假定}$$

5. 结果分析

利用 Python 进行参数估计结果如下：

$$\begin{aligned}\widehat{Y_t} = 226.2288 &- 20.4159X_1 + 0.0319X_2 - 6.7353X_3 - 5.8324X_4 + 1.7482X_5 \\ &+ 11.5864X_6 - 0.1486X_7 - 1.5945X_8 + 4.6523X_9 + 1.6058X_{10} \\ &- 6.0194X_{11}\end{aligned}$$

$$\text{标准差} = (1.889)(0.879)(0.111)(0.443)(0.514)(0.112)(0.702)$$
$$(0.160)(0.215)(0.268)(0.272)(0.130)$$

$$t = (119.78)(-23.224)(0.288)(-15.19)(-11.354)$$
$$(15.653)(16.503)(-0.928)(-7.405)(17.331)(5.9)(-46.299)$$

$$R^2 = 0.557 \quad \bar{R}^2 = 0.556 \quad F = 958.7 \quad df = 8388$$

5.1 模型检验

1. 可决系数和修正可决系数 $R^2 = 0.557 \quad \bar{R}^2 = 0.556$，对于真实模拟数据集而言，模型整体拟合程度尚可，有 55%的因素被回归模型成功解释了。因为可决系数是随解释变量单调不减的函数，影响成绩的因素也有数据集未考虑进去的方面，故想要提高模型的可决系数可以引进更多的解释变量。

2. **F 检验**：针对$H_0: \beta_1 = \beta_2 = \cdots = \beta_{11} = 0$，取$\alpha = 0.05$查自由度为$k - 1 =$

10 和 $n-k=8388$ 的临界值 $F_\alpha(10,8388)=2.54$。由于 $F=958.7>F_\alpha(10,8388)=2.54$, 应拒绝$H_0$。说明在 95%置信水平下回归方程显著，即解释变量$X_1,X_2,\dots,X_{11}$**联合**起来对学生考试总成绩Y_1有显著影响。

3. t **检验**：针对$H_0=0$给定$\alpha=0.05$, 查 t 分布表，在自由度为 8388 时临界值为$t_{0.025}(8388)=1.960$。除了X_2,X_7的$|t|<1.960$以外，其他解释变量的$|t|$均大于 1.960。表示在 5%的显著性水平、其他变量不变的情况下，$X_1,\dots,X_{11}$ (除X_2,X_7) 分别对学生考试总成绩都有显著影响。X_2,X_7的 t 检验不通过，表明可能存在多重共线性。

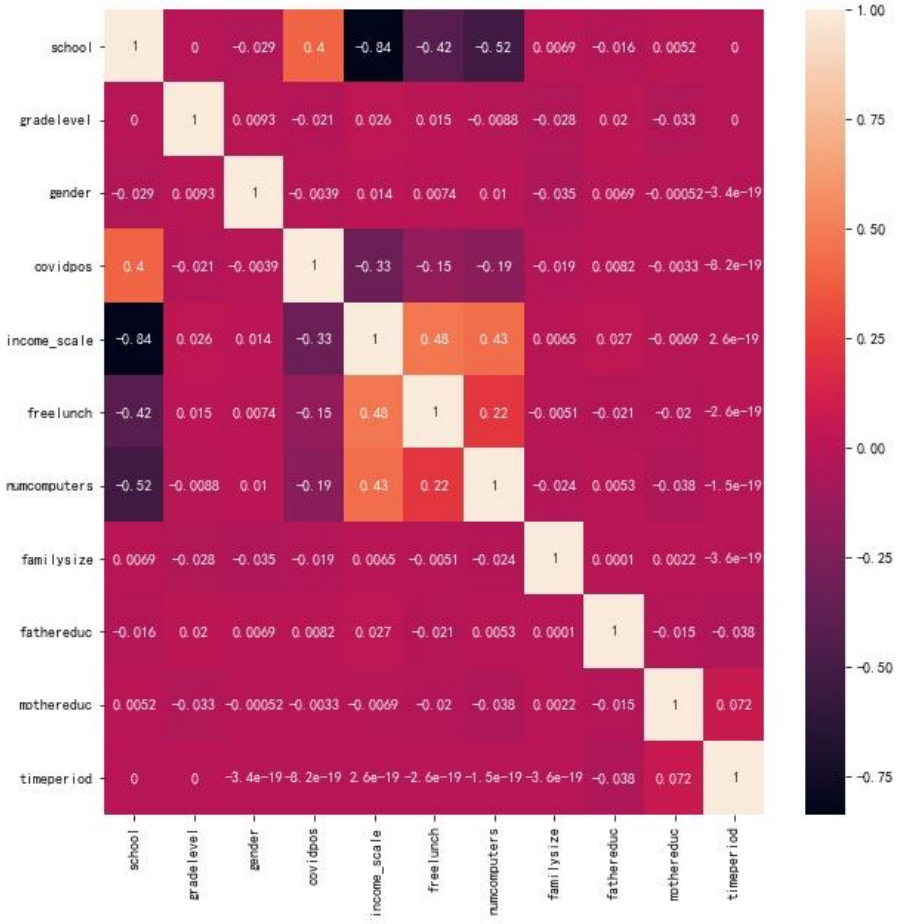
4. 用 P 值检验：X_2,X_7对应的p值 >0.05, 显著性检验不通过。其余解释变量对应的p值 $=0.000<\alpha=0.05$，参数显著性检验通过。

5.2 多重共线性检验

常见的不完全多重共线性会导致参数估计值的方差增大、参数区间估计的置信区间变大、假设检验出现错误判断、可决系数高但参数的 t 检验不显著，甚至估计的系数符号与经验分析相反。

本论文的研究目的不在于精确预测Y值，而在于分析各个解释变量对Y的影响关系，所以要进行多重共线性的检验，如果存在多重共线性，则需进行修正。

在模型分析中，我们发现X_2,X_7的 t 检验不通过，利用简单相关系数检验法，求得X_1,X_5,X_7之间的相关系数$\rho(X_1,X_5)=0.84$，$\rho(X_1,X_7)=0.52$。相关系数图如下所示。这些变量都可以反映家庭的经济条件，所以可以合理认为相互之间存在多重共线性。



使用逐步回归法进行修正：用被解释变量对每一个解释变量做简单回归，以对被解释变量贡献最大的解释变量所对应的回归方程为基础，按对被解释变量贡献大小的顺序逐个引入其余的解释变量。若新变量的引入改进了R^2和 F 检验，且回归参数的 t 检验在统计上也是显著的，则在模型中保留该变量。反之则删除该变量。

对解释变量$X_1,\dots,X_{11}$分别做简单回归

$R^2 = 0.366,0.000,0.009,0.098,0.350,0.152,0.094,0.002,0.021,0.000,0.115$。

故以X_1的简单回归方程为基础，逐步添加变量。

修正后的模型如下，删除了X_2和X_7：

$$\widehat{Y_t} = 226.0163 - 20.1543X_1 - 6.7314X_3 - 5.8467X_4 + 1.7497X_5 + 11.5807X_6 - 1.5916X_8 + 4.6548X_9 + 1.6136X_{10} - 6.0196X_{11}$$

$$\text{标准差} = (1.539)(0.836)(0.443)(0.513)(0.112)(0.702)$$

$$(0.215)(0.268)(0.272)(0.130)$$

$$t = (146.824)(-24.115)(-15.183)(-11.388)$$

$$(15.684)(16.497)(-7.398)(17.344)(5.938)(-46.304)$$

$$R^2 = 0.557 \quad \bar{R}^2 = 0.556 \quad F = 1172 \quad df = 8390$$

该修正改进了模型的 F 检验统计量，各变量 t 检验全部显著，各参数估计值改变不大，无符号变化。

模型意义检验：X_1的系数为-20.15，表示如果该变量从 0（富裕学校）变为 1（贫穷学校），学生的总体成绩将平均下降 20.15 分，说明学校条件对学生成绩的影响非常大，学校财政因素可能制约学生的全面发展、师资聘请、实验条件等。X_3的系数为-6.7，表示如果该变量从 0（女性）变为 1（男性），学生的总体成绩平均会下降 6.7 分，这可能是男女性格不同导致的学习、复习态度差异，从而引起期末成绩差异。X_4的系数为-5.85，表示如果该变量从 0（未得过新冠）变成 1（得过新冠），学生的总体成绩平均会下降 5.85 分，因为生病会影响学习进度和生活质量。X_5的系数为 1.75，表示用在儿童身上的家庭收入每增加 1 万美金，孩子的总体成绩会平均上升 1.75 分，提高孩子的教育和生活条件可以提升学习效果。X_6的系数为 11.58，表示该变量如果从 0（未申请免费餐补）变成 1（申请免费餐补），学生的总体成绩平均会提升 11.58，因为贫困学生会更加珍惜学习机会。X_8的系数为-1.59，表示家庭成员数目每增加 1 人，学生平均成绩会下降 1.59 分，因为家庭人数增多会导致家庭经济负担增大、家庭琐事增多，降低学生的学习条件。X_9的系数为 4.65，X_{10}的系数为 1.61，表示父亲和母亲的学历每增高一个等级，孩子的平均成绩会提高 4.65 分和 1.61 分，这说明了高学历背景的家庭对孩子的学习有辅助提升作用，因为父母更加了解高等教育，可以对孩子进行正确有效的指导。X_{11}的系数为-6.02，表示学期数每增加 1，学生的平均成绩会下降 6.02 分，因为学期数越大，受新冠影响越久，转为线上教学的时间越久，学生越难管控、教学质量越难保证。

5.3 异方差性检验

当异方差性存在时，参数估计值仍然是无偏的，但不是 OLS 估计方差最小的，并且会导致 t 检验失去意义、参数估计失去有效性从而导致 Y 值预测也是无效

的。当存在异方差时，可采用加权最小二乘法进行修正。

使用 Goldfeld-Quanadt 检验的前提条件：检验使用的样本为大样本容量，除了同方差假定不成立外，其他假定均满足。将解释变量的取值从小到大排序并删除中间的$\frac{1}{5}n$个观测值，将剩下的样本分成两部分，构造 F 统计量，提出假设H_0:两部分数据的方差相等。在显著性水平$\alpha = 0.05$的条件下，计算得$F = 0.8 < F_{0.05}(3360,3360) \approx 1$，接受原假设，表示模型中的随机误差不存在异方差。

使用 White 检验的前提条件：不需要关于异方差的任何先验信息，但需要样本为大容量。对 OLS 估计后的残差平方对常数项、解释变量、解释变量的平方及其交叉乘积做辅助回归，计算nR^2，提出假设H_0:辅助回归系数全为 0。在原假设成立的情况下，nR^2服从自由度为 54 的卡方分布。给定显著性水平$\alpha = 0.05$，查卡方分布的临界值，$nR^2 < 101.87 = \chi^2_{0.05}(54)$，接受原假设，表明模型中不存在异方差。

5.4 自相关性检验

在随机误差项存在自相关的条件下，使用 OLS 法将降低估计量的方差、低估真实的σ^2。虽然估计量仍然是无偏的，但不是有效的，即方差不是最小的。因为参数的估计量是无效的，F 检验、t 检验和R^2检验也是不可靠的。模型的预测精度取决于抽样误差和总体误差项的方差σ^2，因为自相关会导致方差估计不可靠，从而导致模型预测值的置信区间不可靠，预测精度降低。

下面使用 DW 检验法和 Breusch-Godfrey 检验（LM 检验）来排查自相关性。

DW 检验前提条件为：解释变量X为非随机的；模型随机误差项为一阶自回归形式；模型的解释变量中不包含滞后的被解释变量；截距项不为 0，即只适用于有常数项的回归模型；数据序列无缺失项。提出原假设$H_0: \rho = 0$，计算得 DW 统计量=1.931，根据样本容量 n=8400，解释变量数目 k=9,查 DW 分布表，得临界值$d_L = 1.675, d_U = 1.863$，因为$d_U = 1.86 < DW = 1.93 < 4 - d_U$，所以误差项不存在自相关。

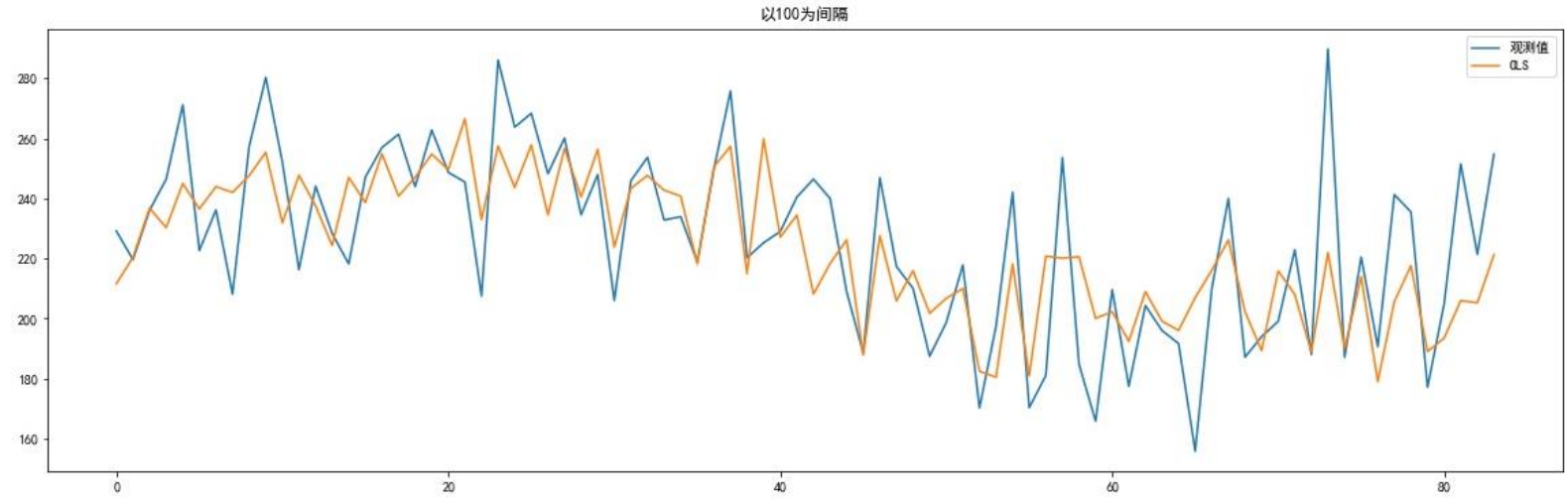
Breusch-Godfrey 检验（LM 检验）的原假设$H_0: \rho_1 = \cdots = \rho_p = 0$，即不存在自相关。用 OLS 估计原模型并得到残差，用残差对解释变量和滞后残差做辅助

回归，计算辅助回归的可决系数，构造统计量$LM = nR^2 = 37.47$，服从卡方分布，p 值 $= 0.4 > 0.05 = \alpha$，表示接受原假设，误差项不存在自相关。

5.5 模型拟合效果可视化

本模型可决系数表示 55%以上的因素被模型成功解释了。因为本文使用的数据集未考虑到的真实世界影响学生学习效果的社会学因素还有很多，而可决系数是随变量个数单调不减的，想要提高可决系数可以增加解释变量，如种族、学生性格、考前是否复习完毕、学生身体素质、心理健康情况等。在社会学定性分析中，可决系数不是唯一衡量模型好坏的标准，不应为了追求高的可决系数胡乱增添不显著的变量，而应更关注模型的逻辑合理性、解释性、显著性从而提出可行的建议。

本文回归模型的拟合值和真实值之间的等间距抽样图像如下所示，蓝色为真实观测值，黄色为 OLS 拟合后的估计值：



可看出模型的拟合值波动与真实值同步，所以对模型的分析能反映真实情况。

6. 结论

本论文旨在分析疫情时代下影响学生学习效果的各方面因素。以学生三门期末考试阅读、写作和数学的成绩之和作为被解释变量，反映学生该学期学习效果。以家庭经济条件、学校条件、线上或线下授课方式、学生性别、父母学历等因素量化作为解释变量，研究它们各自对学习效果的正负向影响和显著性。

本文建模的重点在于各因素定性分析和模型解释，并且给出可行建议，而非精确预测每个学生以后的考试成绩。逐步回归修正多重共线性后的模型 F 检验、t 检验系数显著，正负符号均符合经验判断。通过分析模型系数可得出以下结论：

1.优渥的家庭经济条件和学校条件对学生成绩有显著的促进影响。

2.患过新冠肺炎的学生成绩会显著下降，因为病情会降低学习和生活质量。

3.父母学历越高，孩子的学习效果越好。受过高等教育的父母更了解如何正确引导子女。

4.相比面对面授课，线上教学模式显著降低了学生的学习效果。导致这种现象的主要原因在于远程授课的学生管控难度大，学生自主性难以保持，师生之间的直接交流减少，信息互通不足。

据此，本文建议各教学机构在疫情得到抑制的情况下及时恢复线下教学，提供学生以严格、良好的学习氛围和环境，并保持研究如何提高线上教学的管控，比如随堂点名、及时收取并批改作业、增设定期考试、师生学习进度沟通环节等，以督促学生跟上学习进度；建议各学生注意防范疫情，避免因生病影响正常的学习生活，端正学习态度，珍惜学习机会，认真学习并在考前完成全部的复习计划；建议（准）父母提高自身文化修养，以更科学专业的方式引导子女的学习和生活，保障在孩子教育方面的支出。

引用

[1] 波特兰州立大学 Covid19-effect-on-grades 数据集开源网址
https://www.kaggle.com/dylanbollard/covid19-effect-on-grades-constructed-dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	studentID	school	gradelevel	gender	covidpos	householdincome	freelunch	numcomputers	familysize	fathereduc	mothereduc	readingscore	writingscore	mathscore	readingscoreSL	writingscoreSL	mathscoreSL	timeperiod
2	1	0	6	1	1	59065	0	5	3	1	0	73	69	87	85	83	71	2
3	1	0	6	1	1	59065	0	5	3	1	0	63	73	57	78	75	58	4
4	1	0	6	1	1	59065	0	5	3	1	0	80	67	57	88	80	85	1
5	1	0	6	1	1	59065	0	5	3	1	0	69	56	74	70	48	87	3
6	1	0	6	1	1	59065	0	5	3	1	0	83	87	65	69	86	66	0
7	1	0	6	1	1	59065	0	5	3	1	0	62	37	63	82	78	85	5
8	2	0	6	1	1	107480	1	4	1	1	1	75	91	85	71	59	83	1
9	2	0	6	1	1	107480	1	4	1	1	1	77	82	59	71	81	56	2
10	2	0	6	1	1	107480	1	4	1	1	1	63	77	85	54	73	64	4
11	2	0	6	1	1	107480	1	4	1	1	1	84	60	77	67	55	57	5
12	2	0	6	1	1	107480	1	4	1	1	1	78	81	59	76	71	63	3
13	2	0	6	1	1	107480	1	4	1	1	1	89	87	89	87	73	77	0
14	3	0	6	1	0	79801	1	0	2	1	1	84	85	90	73	78	93	2
15	3	0	6	1	0	79801	1	0	2	1	1	68	72	56	79	61	58	3
16	3	0	6	1	0	79801	1	0	2	1	1	58	84	91	100	86	65	1
17	3	0	6	1	0	79801	1	0	2	1	1	89	68	67	74	74	70	5
18	3	0	6	1	0	79801	1	0	2	1	1	85	66	66	69	68	49	4
19	3	0	6	1	0	79801	1	0	2	1	1	72	88	51	83	80	85	0
20	4	0	6	1	1	125976	1	2	3	1	2	74	93	76	83	78	81	4
21	4	0	6	1	1	125976	1	2	3	1	2	83	55	89	89	89	83	1
22	4	0	6	1	1	125976	1	2	3	1	2	59	73	85	75	89	54	5
23	4	0	6	1	1	125976	1	2	3	1	2	87	92	93	91	88	72	2