SPSS在班级学生数据分析中的应用

摘要

本文通过同济大学经济与管理学院某班的学生数据，使用描述统计、假设检验、方差分析、回归分析等方法分析了性别与身高的关系、学生不同时期的期望月薪关系、不同专业的男女比例、性别与日常花费的关系、不同读研规划的学生的比例、平均绩点和是否读研的独立性等。得出了一系列结论，对于分析学生心理、帮助其合理规划人生道路，有一定的意义。

1 数据来源与说明

本文中的数据来自同济大学经济与管理学院某课程班所有学生填写的问卷。通过问卷获得的变量包括：

变量名称	变量备注
性别	女为0，男为1
年龄	具体数字，单位为岁
身高	具体数字，单位为厘米
年级	具体数字（例如“大一”则为1）
专业	经济金融系为1，管理科学系为2，工商管理系为3
是否计划读研	是为1，否为2，不确定为3
平均绩点	具体数字
期望起始月薪	具体数字，单位为千元
期望工作五年后月薪	具体数字，单位为千元
参加社团数量	具体数字，单位为个
对校园服务的满意度	从不满意（1）到非常满意（4）打分
本学期在教材和日用品方面的花费	具体数字，单位为元

通过收集问卷，共得到70名学生的数据。特别鸣谢管理科学与工程系陈志宗副教授帮助收集、整理数据。

2 数据处理与分析

2.1 描述统计

在描述统计部分，我们将以身高为例进行分析。我们计算所有样本的身高的各统计量并检验其是否近似服从正态分布。

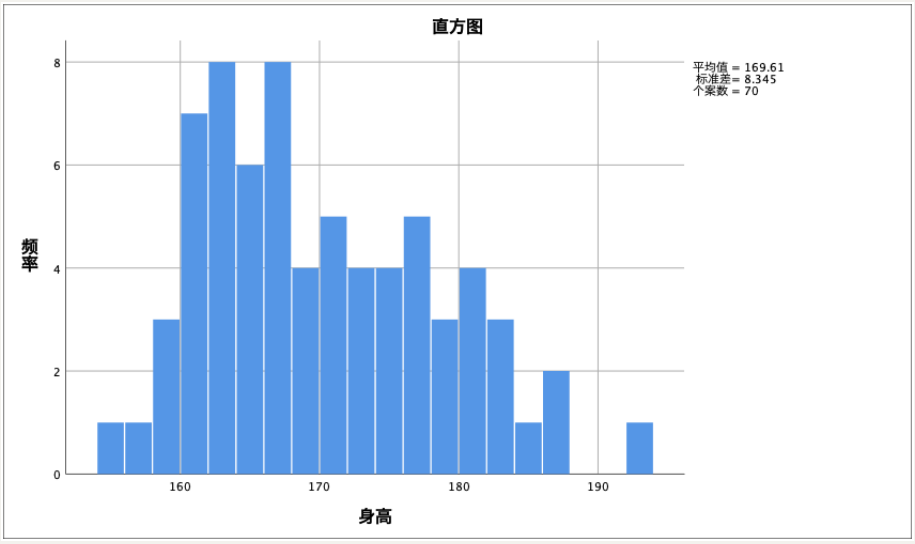
2.1.1 身高数据的各统计量计算

```
EXAMINE VARIABLES=height
  /PLOT BOXPLOT STEMLEAF HISTOGRAM NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

计算得到身高变量的各统计量，并画出直方图：

描述

			统计	标准误差
身高	平均值		169.61	0.997
	平均值的 95% 置信区间	下限	167.62	
		上限	171.60	
	5% 剪除后平均值		169.31	
	中位数		168.00	
	方差		69.632	
	标准偏差		8.345	
	最小值		155	
	最大值		193	
	范围		38	
	四分位距		13	
	偏度		0.519	0.287
	峰度		-0.387	0.566



班级学生的身高主要分布在162-176厘米处，身高均值的95%水平置信区间为167.62-171.60厘米；从偏度系数$\beta_S = 0.519$以及直方图中可以看出，全班学生身高的分布是一个右偏分布。

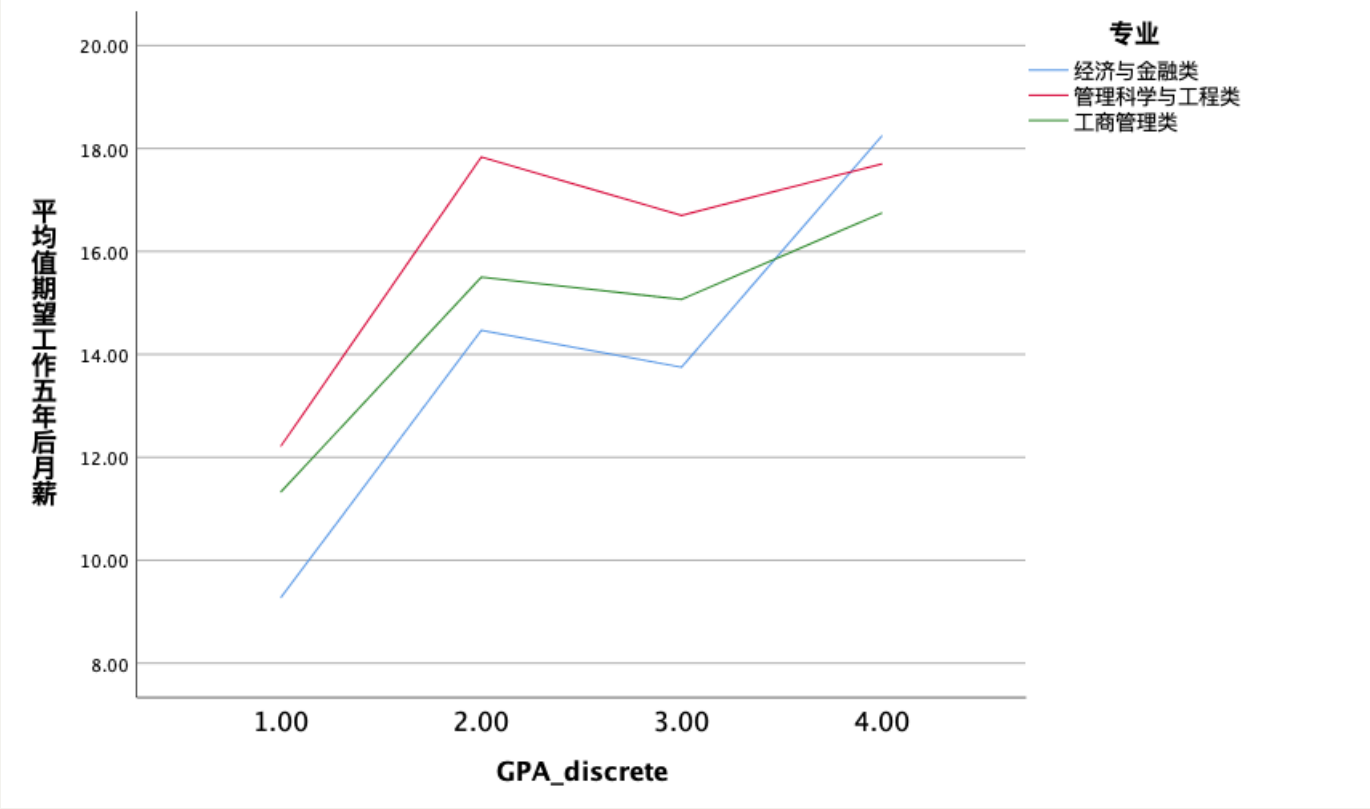
2.1.2 不同绩点、不同专业期望月薪的描述统计

首先，根据平均绩点的分位数对平均绩点进行离散化，得到`GPA_discrete`变量。

```
RECODE GPA (0 thru 3.8=1) (3.8 thru 4.17=2) (4.17 thru 4.4=3) (4.4
thru 5=4) INTO GPA_discrete.
EXECUTE.
```

接着作出折线图：

```
GRAPH
  /LINE(MULTIPLE)=MEAN(salaryAfter5Year) BY GPA_discrete BY major.
```



可以直观的看出，期望工作五年后月薪与平均绩点有一定的关系；整体来看，管理科学与工程类的学生的期望月薪比其他两个专业高。

2.2 假设检验

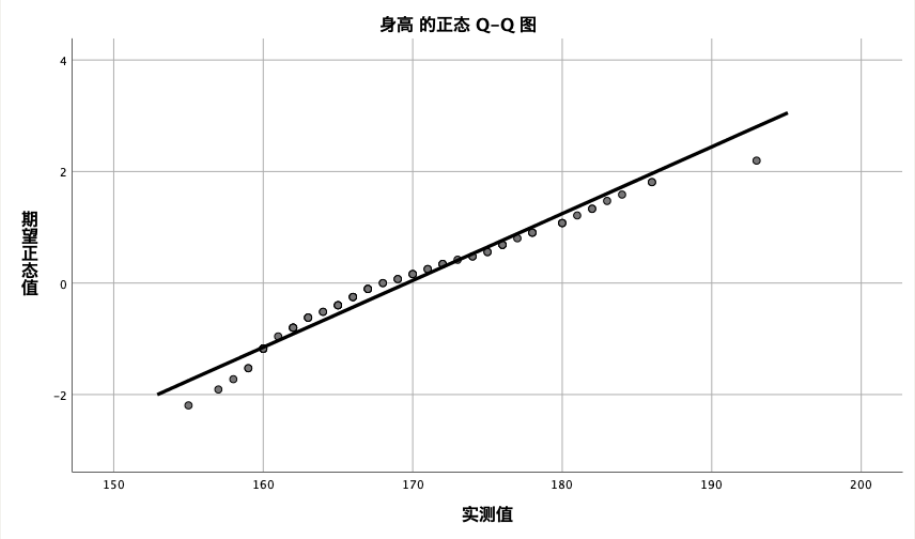
2.2.1 身高数据的正态性检验

通过对全班学生的身高进行正态性检验，得到结果：

正态性检验

	柯尔莫戈洛夫-斯米诺夫(V)[a]			夏皮洛-威尔克		
	统计	自由度	显著性	统计	自由度	显著性
身高	0.109	70	0.039	0.964	70	0.040

a. 里利氏显著性修正

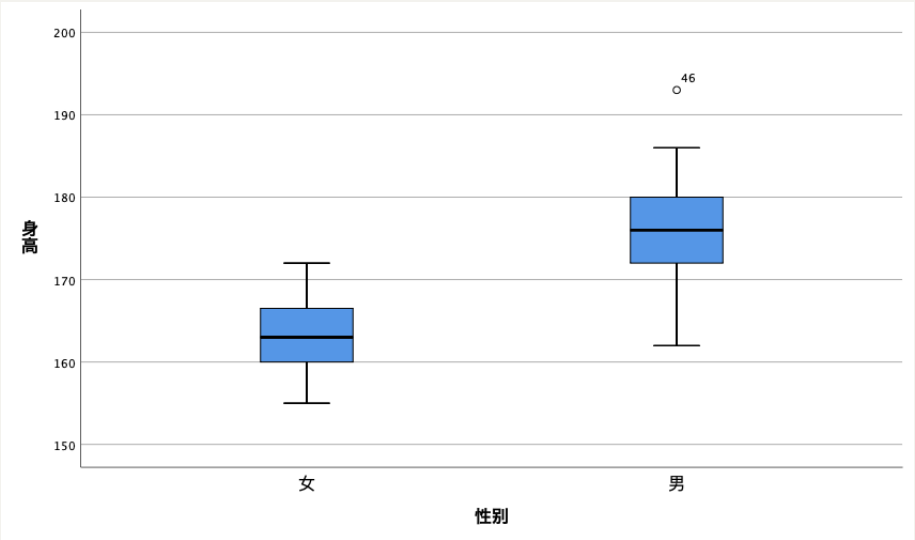


全班学生身高的正态性不显著。我们尝试将全班学生按性别分为两类。

```
SORT CASES  BY sex.
SPLIT FILE SEPARATE BY sex.
```

通过箱线图，可以看出不同性别学生之间身高的差异：

```
EXAMINE VARIABLES=height BY sex
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL.
```



对两组样本进行正态性检验，得到结果：

正态性检验[a]

		柯尔莫戈洛夫-斯米诺夫(V)[b]			夏皮洛-威尔克		
		统计	自由度	显著性	统计	自由度	显著性
身高		0.112	36	.200*	0.984	36	0.857

*. 这是真显著性的下限。

a. 性别 = 女

b. 里利氏显著性修正

正态性检验[a]

		柯尔莫戈洛夫-斯米诺夫(V)[b]			夏皮洛-威尔克		
		统计	自由度	显著性	统计	自由度	显著性
身高		0.070	34	.200*	0.992	34	0.995

*. 这是真显著性的下限。

a. 性别 = 男

b. 里利氏显著性修正

按性别分类的两组学生身高数据均通过正态性检验。

2.2.2 平均绩点数据的正态性检验

同样，全班学生的平均绩点不通过正态性检验；在按专业分类后，经济与金融类的学生平均绩点正态性显著性不高，其他专业则通过正态性检验：

正态性检验

		柯尔莫戈洛夫-斯米诺夫(V)[a]			夏皮洛-威尔克		
专业		统计	自由度	显著性	统计	自由度	显著性
平均绩点	经济与金融类	0.176	28	0.026	0.875	28	0.003
	管理科学与工程类	0.143	24	.200*	0.958	24	0.397
	工商管理类	0.117	18	.200*	0.966	18	0.725

*. 这是真显著性的下限。

a. 里利氏显著性修正

2.2.3 性别对身高的影响显著性检验

```
T-TEST GROUPS=sex(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=height
  /CRITERIA=CI(.95).
```

通过对不同性别样本的身高数据进行独立样本t检验：

$$H_0$$：男、女学生身高均值相等 vs H_1：男、女学生身高均值不相等

得到以下结果：

独立样本检验

		莱文方差等同性检验		平均值等同性 t 检验						
									差值 95% 置信区间	
		F	显著性	t	自由度	Sig.（双尾）	平均值差值	标准误差差值	下限	上限
身高	假定等方差	5.079	0.027	-9.944	68	0.000	-12.760	1.283	-15.320	-10.199
	不假定等方差			-9.810	53.593	0.000	-12.760	1.301	-15.368	-10.152

显著性明显小于$\alpha = 0.05$，因此拒绝原假设，男女学生身高具有显著性差异。性别对身高的影响显著。

2.2.4 期望起始月薪与工作五年后月薪相关性检验

```
T-TEST PAIRS=expectedSalary WITH salaryAfter5Year (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

通过对期望起始月薪与工作五年后月薪数据进行相关性检验：

$$H_0$$：两期望月薪相关系数为0 vs H_1：两期望月薪相关系数不为0

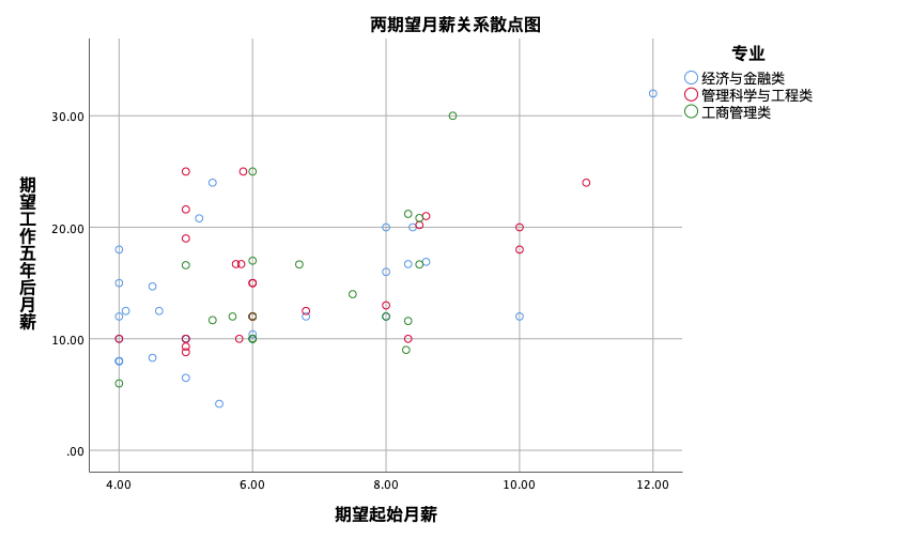
得到以下结果：

配对样本相关性

		N	相关性	显著性
配对 1	期望起始月薪 & 期望工作五年后月薪	70	0.486	0.000

根据相关系数$\rho = 0.486$，显著性明显小于$\alpha = 0.05$，因此拒绝原假设，两期望月薪之间具有一定的相关性。

作出散点图，可以直观地看出两期望月薪之间的正相关关系。



2.2.5 不同专业男女比例的检验

```
SPLIT FILE SEPARATE BY major.
NPAR TESTS
  /BINOMIAL (0.50)=sex
  /MISSING ANALYSIS.
```

将所有数据按专业进行划分，进行二项检验：

$$H_0：女性占总体比例为0.5 \text{ vs } H_1：女性占总体比例不为0.5$$

结果如下：

二项检验[a]

		类别	个案数	实测比例	检验比例	精确显著性（双尾）
性别	组 1	女	17	0.61	0.50	0.345
	组 2	男	11	0.39		
	总计		28	1.00		

a. 专业 = 经济与金融类

二项检验[a]

		类别	个案数	实测比例	检验比例	精确显著性（双尾）
性别	组 1	女	10	0.42	0.50	0.541
	组 2	男	14	0.58		
	总计		24	1.00		

a. 专业 = 管理科学与工程类

二项检验[a]

		类别	个案数	实测比例	检验比例	精确显著性（双尾）
性别	组 1	女	9	0.50	0.50	1.000
	组 2	男	9	0.50		
	总计		18	1.00		

a. 专业 = 工商管理类

显著性表明，对于三个专业的该检验，均接受原假设。因此，三个专业的男女比例都较为均衡。

2.2.6 性别与日常花费间关系的游程检验

首先，将数据按日常花费变量进行排序：

```
SORT CASES BY cost(D).
```

再通过排序完成的数据对性别进行随机游程检验：

$$H_0：性别排列随机 \text{ vs } H_1：性别排列不随机$$

```
SORT CASES BY cost(D).
NPAR TESTS
  /RUNS(0.5)=sex
  /MISSING ANALYSIS.
```

游程检验

	性别
检验值[a]	0.50
总个案数	70
游程数	31
Z	-1.198
渐近显著性（双尾）	0.231

a. 由用户指定。

显著性较高，接受原假设。可以认为性别与日常花费间关系不大。

如通过独立样本t检验，也可得到相同的结论：

独立样本检验

		莱文方差等同性检验		平均值等同性 t 检验						
									差值 95% 置信区间	
		F	显著性	t	自由度	Sig.（双尾）	平均值差值	标准误差差值	下限	上限
购买教材与日用品花费	假定等方差	1.722	0.194	1.462	68	0.148	189.565	129.690	-69.228	448.358
	不假定等方差			1.467	67.750	0.147	189.565	129.246	-68.358	447.489

2.2.7 不同读研规划学生之间比例关系的检验

建立如下假设：

H_0：读研规划为$(1,2,3)$的学生比例为$5:1:2$ vs H_1：该比例不满足

其中，计划读研为1，不计划读研为2，不确定为3。进行χ^2检验：

```
NPAR TESTS
  /CHISQUARE=graduate
  /EXPECTED=5 1 2
  /MISSING ANALYSIS.
```

得到如下结果：

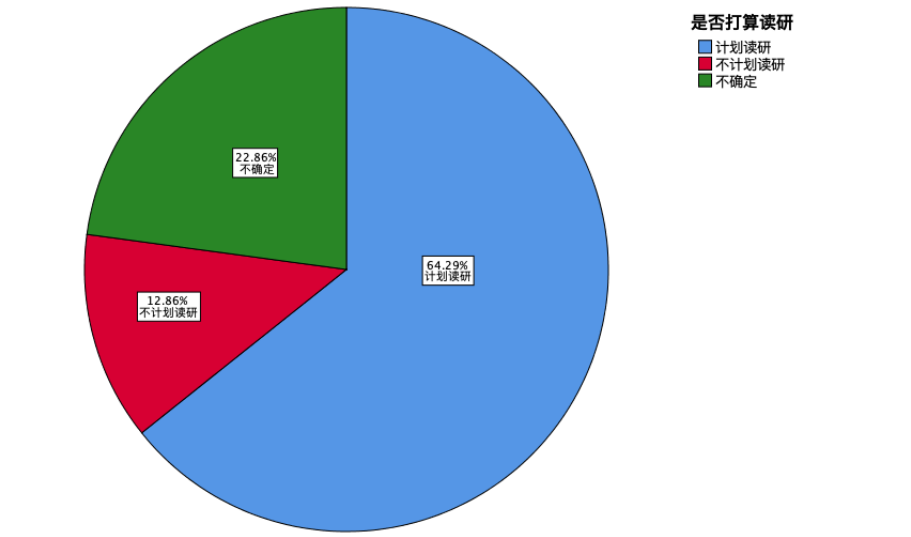
检验统计

	是否打算读研
卡方	.171[a]
自由度	2
渐近显著性	0.918

a. 0 个单元格 (0.0%) 的期望频率低于 5。期望的最低单元格频率为 8.8。

显著性很高，接受原假设。不同读研规划学生之间的比例满足$5:1:2$。

通过饼图，可以直观地看出这一点：



2.2.8 是否计划读研与平均绩点间的独立性检验

按照是否读研和平均绩点的离散分组进行统计，得到交叉表：

```
DATASET DECLARE graduate_GPA.
AGGREGATE
  /OUTFILE='graduate_GPA'
  /BREAK=graduate GPA_discrete
  /N_BREAK=N.
```

再对交叉表加权后进行χ^2检验：

$$H_0$$：两变量之间相互独立 vs H_1：两变量之间不相互独立

```
CROSSTABS
  /TABLES=graduate BY GPA_discrete
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

卡方检验

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	12.565[a]	6	0.050
似然比	14.742	6	0.022
线性关联	3.385	1	0.066
有效个案数	70		

a. 8 个单元格 (66.7%) 的期望计数小于 5。最小期望计数为 2.06 。

显著性略大于0.05，可在$\alpha = 0.05$水平下接受原假设。

2.3 方差分析

2.3.1 平均绩点因素对期望月薪与日常花费的影响分析

进行单因素方差分析，建立假设：

H_0：各绩点水平学生的期望月薪/日常花费均值相等

vs

H_1：各绩点水平学生的期望月薪/日常花费均值不完全相等

```
ONEWAY expectedSalary salaryAfter5Year cost BY GPA_discrete
  /MISSING ANALYSIS
  /POSTHOC=SNK ALPHA(0.05).
```

ANOVA

		平方和	自由度	均方	F	显著性
期望起始月薪	组间	44.792	3	14.931	4.727	0.005
	组内	208.450	66	3.158		
	总计	253.242	69			
期望工作五年后月薪	组间	409.139	3	136.380	4.775	0.005
	组内	1884.958	66	28.560		
	总计	2294.097	69			
购买教材与日用品花费	组间	987688.359	3	329229.453	1.106	0.353
	组内	19639610.727	66	297569.859		
	总计	20627299.086	69			

根据结果，对于期望月薪，显著性较低，因此拒绝原假设，绩点水平对学生的期望月薪有影响；对于日常花费，则接受原假设，绩点水平对学生的日常花费影响显著性不高。通过S-N-K事后检验，我们可以看出不同绩点水平学生的期望月薪的差异：

期望起始月薪

S-N-K[a,b]

GPA_discrete	个案数	Alpha 的子集 = 0.05 1	2
1.00	18	5.4422	
3.00	19	6.0842	
2.00	17	6.6153	6.6153
4.00	16		7.6650
显著性		0.133	0.086

将显示齐性子集中各个组的平均值。

a. 使用调和平均值样本大小 = 17.428。

b. 组大小不相等。使用了组大小的调和平均值。无法保证 I 类误差级别。

期望工作五年后月薪

S-N-K[a,b]

GPA_discrete	个案数	Alpha 的子集 = 0.05 1	2
1.00	18	10.9317	
3.00	19		14.8037
2.00	17		15.9588
4.00	16		17.5106
显著性		1.000	0.300

将显示齐性子集中各个组的平均值。

a. 使用调和平均值样本大小 = 17.428。

b. 组大小不相等。使用了组大小的调和平均值。无法保证 I 类误差级别。

2.4 回归分析

2.4.1 变量之间的相关性分析

尝试计算平均绩点、期望起始月薪、期望工作五年后月薪、购买教材与日用品花费等变量之间的相关系数。

```
CORRELATIONS
  /VARIABLES=GPA expectedSalary salaryAfter5Year cost
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

相关性

		平均绩点	期望起始月薪	期望工作五年后月薪	购买教材与日用品花费
平均绩点	皮尔逊相关性	1	.330**	.365**	0.090
	Sig.（双尾）		0.005	0.002	0.457
	个案数	70	70	70	70
期望起始月薪	皮尔逊相关性	.330**	1	.486**	-0.105
	Sig.（双尾）	0.005		0.000	0.385
	个案数	70	70	70	70
期望工作五年后月薪	皮尔逊相关性	.365**	.486**	1	0.074
	Sig.（双尾）	0.002	0.000		0.544
	个案数	70	70	70	70
购买教材与日用品花费	皮尔逊相关性	0.090	-0.105	0.074	1
	Sig.（双尾）	0.457	0.385	0.544	
	个案数	70	70	70	70

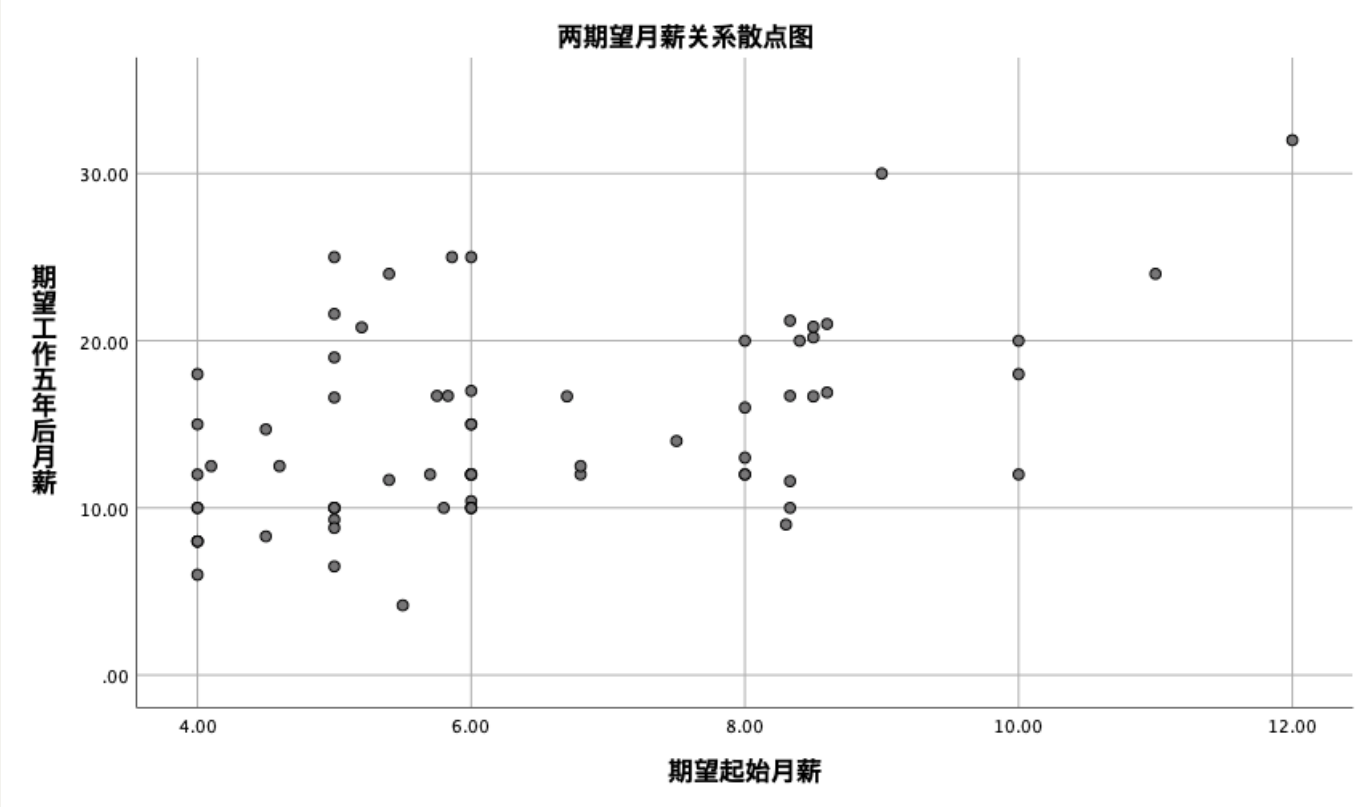
**. 在 0.01 级别（双尾），相关性显著。

可以看出，平均绩点、期望起始月薪、期望工作五年后月薪三个变量之间的相关性较大。

2.4.2 两期望月薪间的回归分析

首先做出期望起始月薪、期望工作五年后月薪的散点图：

```
GRAPH
  /SCATTERPLOT(BIVAR)=expectedSalary WITH salaryAfter5Year
  /MISSING=LISTWISE
  /TITLE='两期望月薪关系散点图'.
```



以期望起始月薪(x)为自变量，期望工作五年后月薪(y)为因变量，进行一元线性回归计算：

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT salaryAfter5Year
  /METHOD=ENTER expectedSalary
  /SCATTERPLOT=(*ZPRED ,*ZRESID) (*ZRESID ,*ZRESID)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

得到如下结果：

系数[a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	5.340	2.133		2.503	0.015
	期望起始月薪	1.461	0.319	0.486	4.580	0.000

a. 因变量：期望工作五年后月薪

从而可以写出回归方程$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}$，其中$\hat{\beta}_0 = 5.340, \hat{\beta}_1 = 1.461$。通过$t$检验：

$$H_0 : \hat{\beta}_1 = 0 \text{ vs } H_1 : \hat{\beta}_1 \neq 0$$

显著性很小，因此拒绝原假设，回归有效。尽管如此，经计算R^2如下：

模型摘要[b]

模型	R	R 方	调整后 R 方	标准估算的错误
1	.486[a]	0.236	0.225	5.07769

a. 预测变量：(常量), 期望起始月薪

b. 因变量：期望工作五年后月薪

R^2较小，说明用线性回归进行拟合效果并不优良。因此，考虑非线形回归，同时将男性与女性分开进行回归：

```
SORT CASES  BY sex.
SPLIT FILE SEPARATE BY sex.

* Curve Estimation.
TSET NEWVAR=NONE.
CURVEFIT
  /VARIABLES=salaryAfter5Year WITH expectedSalary
  /CONSTANT
  /MODEL=LINEAR LOGARITHMIC QUADRATIC CUBIC EXPONENTIAL
  /PLOT FIT.
```

得到以下结果：

模型摘要和参数估算值[a]

因变量:

	模型摘要					参数估算值			
方程	R 方	F	自由度 1	自由度 2	显著性	常量	b1	b2	b3
线性	0.387	21.505	1	34	0.000	3.868	1.657		
对数	0.350	18.306	1	34	0.000	-4.395	10.473		
二次	0.419	11.897	2	33	0.000	13.382	-1.265	0.203	
三次	0.443	8.477	3	32	0.000	-14.649	11.420	-1.571	0.077
指数	0.342	17.702	1	34	0.000	6.570	0.111		

自变量为 期望起始月薪。

a. 性别 = 女

模型摘要和参数估算值[a]

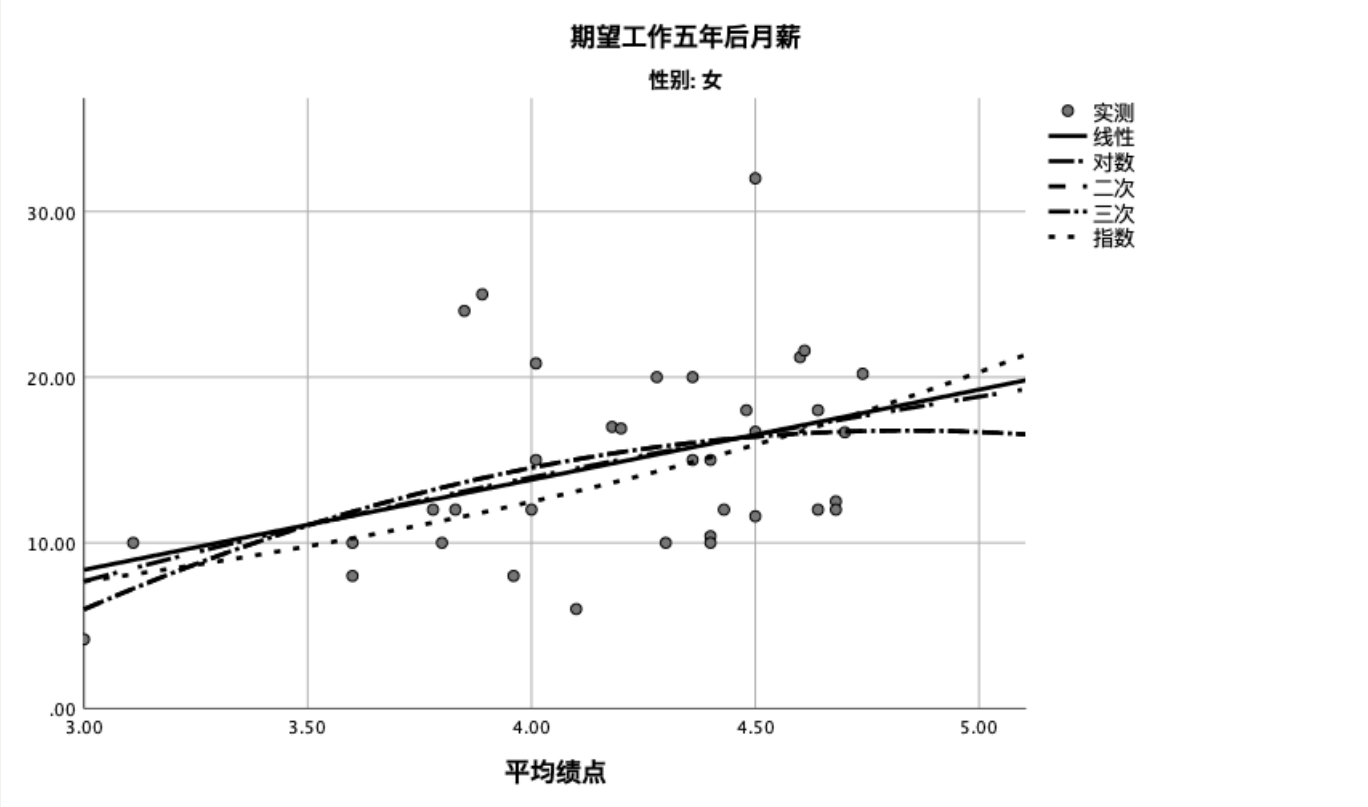
因变量:

	模型摘要					参数估算值			
方程	R 方	F	自由度 1	自由度 2	显著性	常量	b1	b2	b3
线性	0.084	2.932	1	32	0.097	7.900	1.075		
对数	0.082	2.862	1	32	0.100	2.224	6.873		
二次	0.088	1.487	2	31	0.242	14.259	-0.925	0.148	
三次	0.142	1.659	3	30	0.197	-70.854	39.140	-5.914	0.295
指数	0.092	3.230	1	32	0.082	8.595	0.074		

自变量为 期望起始月薪。

a. 性别 = 男

可以看出，男性学生对于采用的五种方程，两期望薪酬之间不具有显著的回归关系；对于女性学生，则具有显著的回归关系，且用三次方程进行拟回归效果最佳。对于女性学生的回归结果如下：



因此，对于女性学生，计算出期望起始月薪的平方值与次方值，使用“步进法”进行线形回归计算：

```
COMPUTE expectedSalarySquare=expectedSalary ** 2.
EXECUTE.
COMPUTE expectedSalaryPower=expectedSalary ** 3.
EXECUTE.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT salaryAfter5Year
  /METHOD=STEPWISE expectedSalary expectedSalarySquare
expectedSalaryPower
  /SCATTERPLOT=(*ZPRED ,*ZRESID) (*ZRESID ,*ZRESID)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

得到如下结果：

模型摘要[a,c]

模型	R	R 方	调整后 R 方	标准估算的错误
1	.651[b]	0.424	0.407	4.52317

a. 性别 = 女

b. 预测变量：(常量), expectedSalaryPower

c. 因变量：期望工作五年后月薪

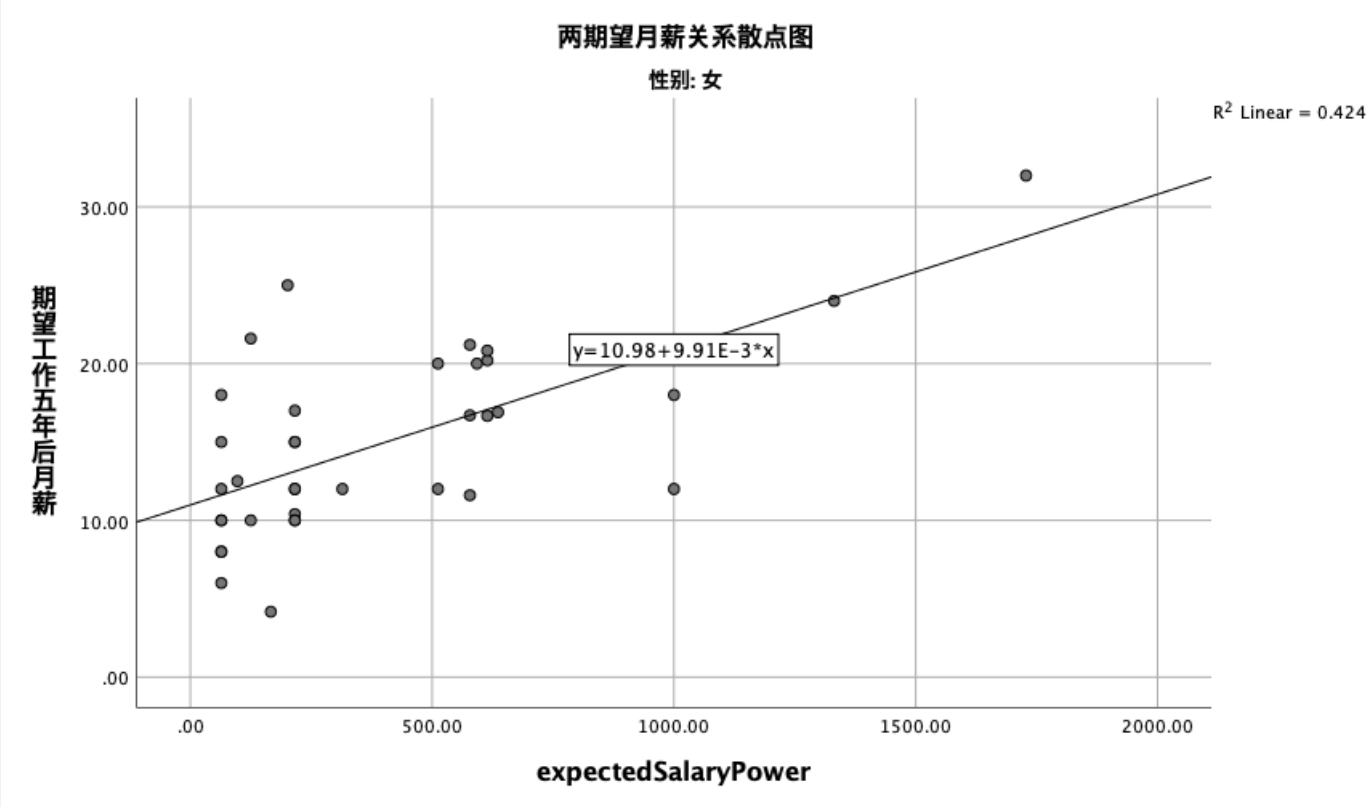
系数[a,b]

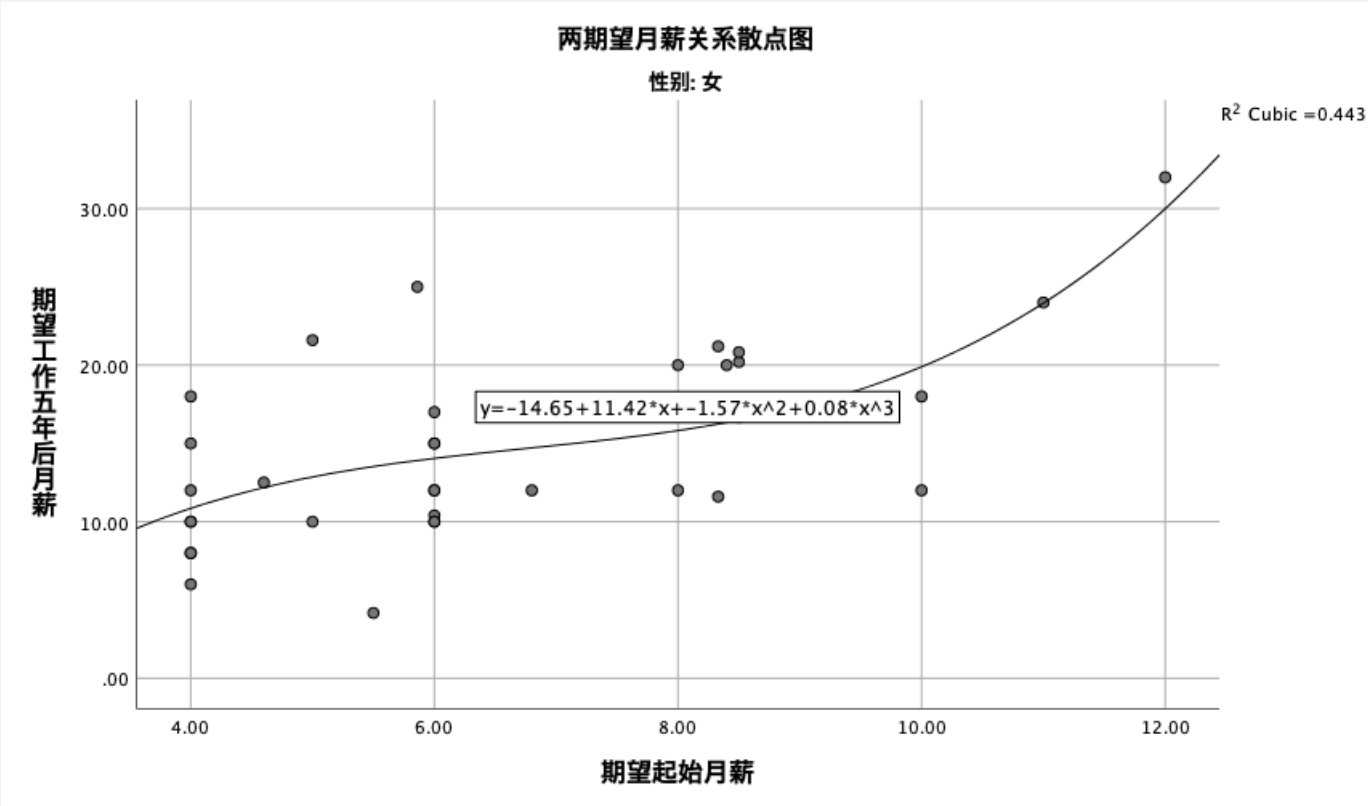
模型		未标准化系数		标准化系数		
		B	标准错误	Beta	t	显著性
1	(常量)	10.985	1.084		10.135	0.000
	expectedSalaryPower	0.010	0.002	0.651	5.005	0.000

a. 性别 = 女

b. 因变量：期望工作五年后月薪

显著性较小，因此回归有效。通过作图，可以直观看出回归效果：





记一给定x数据的女性的期望工作五年后月薪为y_0，下对$E(y_0), y_0$进行估计，其中$x = 8.5$:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10) CIN(95)
  /NOORIGIN
  /DEPENDENT salaryAfter5Year
  /METHOD=STEPWISE expectedSalaryPower
  /SCATTERPLOT=(*ZPRED ,*ZRESID) (*ZRESID ,*ZRESID)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /SAVE PRED MCIN ICIN.
```

得到结果为：

预测结果	$E(y_0)$上限	$E(y_0)$下限	y_0上限	y_0下限
17.1	15.3	18.8	7.7	26.4

2.4.3 平均绩点与期望薪酬间的回归分析

以期望起始月薪(x_1)与平均绩点(x_2)为自变量，期望工作五年后月薪(y)为因变量，使用“步进法”进行多元线性回归计算：

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT salaryAfter5Year
  /METHOD=STEPWISE GPA expectedSalary
  /SCATTERPLOT=(*ZPRED ,*ZRESID) (*ZRESID ,*ZRESID)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

得到结果如下：

系数[a,b]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	3.868	2.499		1.548	0.131
	期望起始月薪	1.657	0.357	0.622	4.637	0.000

a. 性别 = 女

b. 因变量：期望工作五年后月薪

系数[a,b]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	(常量)	-2.842	8.343		-0.341	0.736
	平均绩点	4.436	2.118	0.347	2.095	0.044

a. 性别 = 男

b. 因变量：期望工作五年后月薪

根据结果，女性的期望工作五年后月薪与期望起始月薪具有较显著的回归关系，与平均绩点的回归关系不显著；男性学生则相反。

3 总结

通过选取班级学生的相关数据进行分析，我感受到统计的魅力。统计学是看待世界的另一视角，通过统计得出的结论，有些是符合常识的，而有些是反常识的，不禁让我们去反思，是我们的常识出了问题，还是运用的统计模型不匹配。人的价值观、世界观，也在这样的矛盾中，不断发展。