

第三讲：均值比较与回归分析

教学目的：能应用 SPSS 软件进行：单个均值假设检验、均值比较分析、相关分析、回归分析等

教学内容：1) 均值的比较
2) 相关分析
3) 回归分析

教学重点：均值比较分析、回归分析

教学难点：均值比较分析

教学时间：1 学时

均值的比较 Compare Means

调查研究中的个案 (Cases) 被称为样本。如果样本来自总体，那么，总体的特征可以采用集中趋势或离中趋势加以描述和统计，其结果可以准确地描述总体。一般地，数据总体的均值应为 0，方差应为 1，即服从标准正态分布。现实中，样本的均值与方差都不能满足该条件，但可加大样本规模使之分布接近总体的正态分布。

在 SPSS 中，将两个总体均值近比较称为 Compare Means，可选择 Analyze→Compare Means 来实现。Compare Means 集中了几个用于计量资料均值间比较的过程。具体有：

- Means 过程：对准备比较的各组计算描述指标，进行预分析，也可直接比较。
- One-Samples T Test 过程：进行样本均值与已知总体均值的比较。
- Independent-Samples T Test 过程：进行两样本均值差别的比较，即通常所说的两组资料的 t 检验。
- Paired-Samples T Test 过程：进行配对资料的显著性检验，即配对 t 检验。
- One-Way ANOVA 过程：进行两组及多组样本均值的比较，即成组设计的方差分析，还可进行随后的两两比较。

1.1 Means 过程

和上一章所讲述的几个专门的描述过程相比，Means 过程的优势在于各组的描述指标被放在一起便于相互比较，并且如果需要，可以直接输出比较结果，无须再次调用其他过程。

显然要方便得多。

1.1.1 界面说明

选择 Analyze→Compare Means→Means，进入 Means 对话框，见图 1.1a。其各部分解释如下：

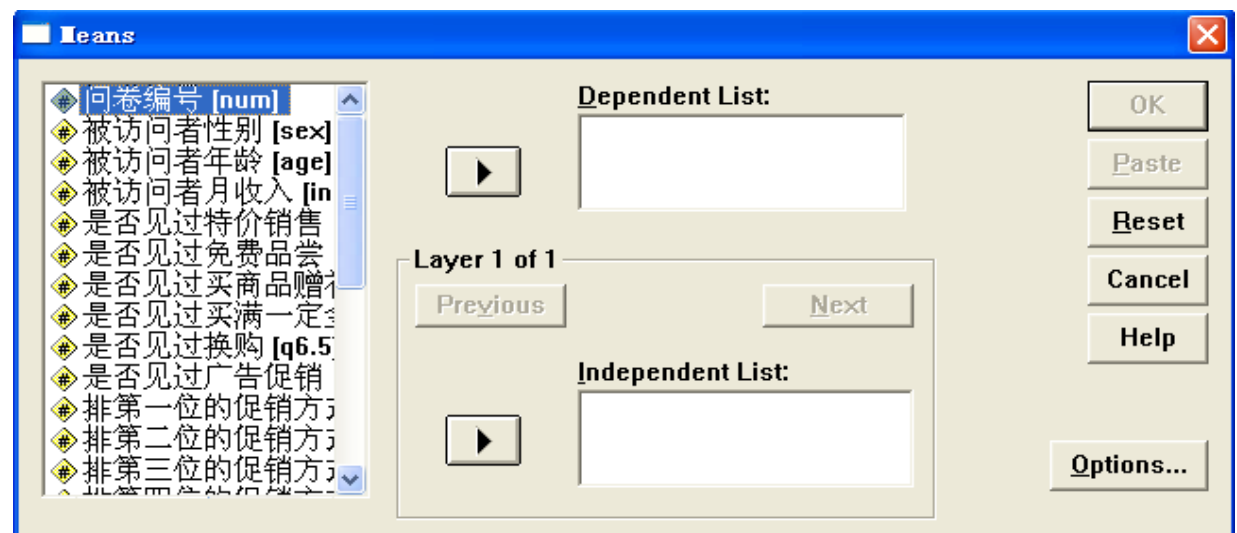


图 1.1a Means 对话框

- Dependent List 框：用于选入需要分析的变量。
- Independent List 框：用于选入分组变量。
- Options：弹出 Options 对话框（见图 1.1b），选择需要计算的描述统计量和统计分析：

- Statistics 框：可选的描述统计量。它们是：

1. sum, number of cases 总和，记录数
2. mean, geometric mean, harmonic mean 均值，几何均值，修正均值
3. standard deviation, variance, standard error of the mean 标准差，均值的标准误，方差
4. median, grouped median 中位数，频数表资料中位数（比如 30 岁组有 5 人，40 岁组有 1 人，则在计算 grouped median 时均按组中值 35 和 45 进行计算）。
5. minimum, maximum, range 最小值，最大值，全距
6. kurtosis, standard error of kurtosis 峰度系数，峰度系数的标准误
7. skewness, standard error of skewness 偏度系数，偏度系数的标准误
8. percentage of total sum, percentage of total N 总和的百分比，样本例数的百分比

- Cell Statistics 框：选入的描述统计量。
- Statistics for First layer 复选框组

1. Anova table and eta 对分组变量进行单因素方差分析，并计算用于度量变量相关程度的 eta 值。
2. Test for linearity 检验线性相关性，实际上就是上面的单因素方差分析。

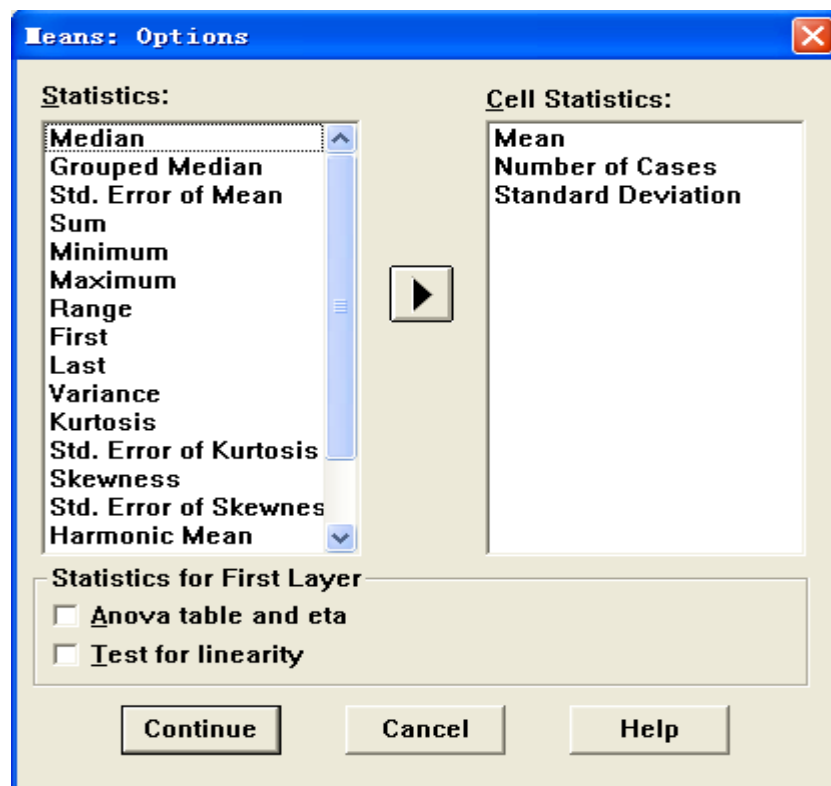


图 1.1b Means 中的 Options 对话框

1.1.2 分析案例

例 1.1 利用 111.sav 文件中的数据分析，不同性别 sex、月收入 income、年龄 age 等 q9（即被访问者最近一次参加促销活动的消费）的不同表现。

上述问题采用 Means 来解决。如果分析消费与性别的关系，或者说研究男女消费的差异，则月收入 and 年龄就是两个控制变量。当然，也可分析消费与与收入的关系、消费与年龄的关系，相应地，另两个变量就成了控制变量了。

这里只给出男女消费差异求解的简化操作：

1. Analyze→Compare Means→Means
2. Dependent list 框：选入 q9
3. Independent list 框：依次选入 sex、income、age（注意：sex 一定要放在第一位）
4. 单击 option：选中 Anova table and eta 复选框，单击 Continue
5. 单击 OK

1.1.3 结果解释

有了上一章的基础，Means 过程的输出看起来就不太困难了。它的输出结果包括 Case

Processing Summary、Report、ANOVA Table、Measures of Association 等。

●缺失值报告。312 个 Cases 均有效。

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
被访问者最近一次参加促销活动的消费 * 被访问者性别 * 被访问者年龄 * 被访问者月收入	312	100.0%	0	.0%	312	100.0%

●常用统计描述量报表。由于 Report 表太长，这里只给出了一部分，但人可以看出表的结构。表中的结果是按默认情况输出均值、样本量和标准差。因为选择了分组变量，所以三项指标均给出分组及合计值，可见以这种方式列出统计量可以非常直观的进行各组间的比较。

Report

被访问者最近一次参加促销活动的消费					
被访问者性别	被访问者年龄	被访问者月收入	Mean	N	Std. Deviation
女	25岁以下	1000元以下	154.60	20	60.309
		1000-1500元	203.13	8	53.850
		1500-2000元	151.00	2	77.782
		Total	167.30	30	61.443
	25-35岁	1000元以下	171.41	51	68.402
		1000-1500元	146.87	31	64.101

●单因素方差分析表。在选择了 Anova table and eta 或 Test for linearity 复选框时出现。实际上就是在检验各组间均值有无差异。表中结果显示：组间 Between Groups 的离差平方和为 154.914，自由度为 1（即只有一个因素 Sex）；而组内 Within Groups 的离差平方和为 1415811.715，自由度为 310；最后 F 值为 0.033，F 值的概率为 0.851>>0.05，表明没有理由拒绝系统默认的原假设——不同性别的消费相同，可认为男女参加促销活动的消费没有什么区别。

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
被访问者最近一次参加促销活动的消费 * 被访问者性别	Between Groups (Combined)	154.914	1	154.914	.033	.851
	Within Groups	1415811.715	310	4728.102		
	Total	1411021.179	311			

- 相关性度量指标，给出 Eta 值以及 Eta 值的平方根。表中数据说明两者关系较弱。

Measures of Association

	Eta	Eta Squared
被访问者最近一次参加促销活动的消费 * 被访问者性别	.010	.000

1.2 One-Samples T Test 过程

One-Samples T Test 过程用于进行样本所在总体均值与已知总体均值的比较，可以自行定义已知总体均值为任意值，该对话框的界面非常简单。

1.2.1 界面说明

选择 Analyze→Compare Means→One-Samples T Test，进入对话框，见图 1.2a。其各部分解释如下：

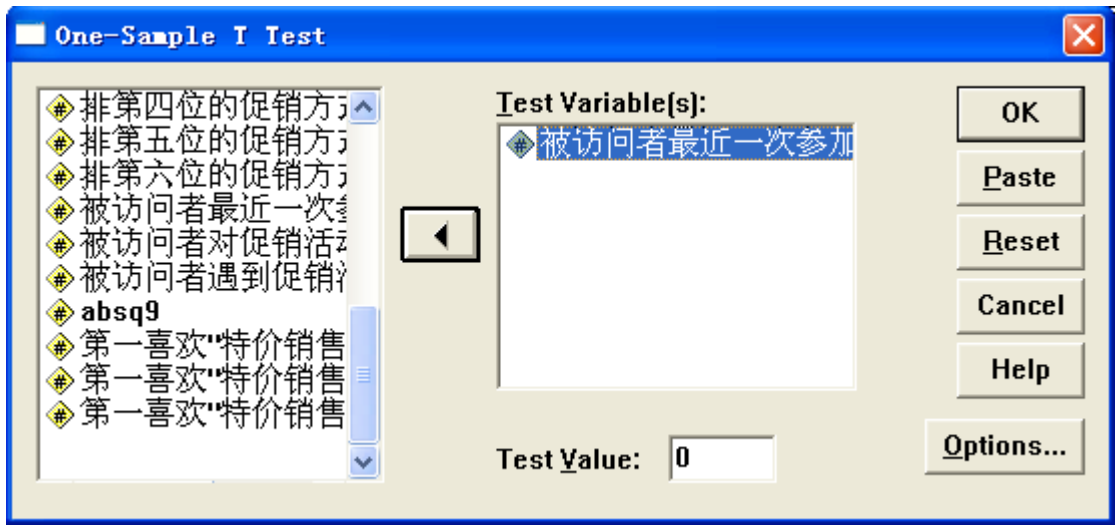


图 1.2a One-Samples T Test 对话框

- Test Variables 框：用于选入需要分析的变量。
- Test Value 框：在此处输入已知的总体均值，默认值为 0。
- Options：弹出 Options 对话框（见图 1.2b），用于定义相关的选项，有：
 - Confidence Interval 框 输入需要计算的均值差值可信区间范围，默认为 95%。如果是和总体均值为 0 相比，则此处计算的就是样本所在总体均值的可信区间。

- Missing Values 单选框组 定义分析中对缺失值的处理方法，可以是具体分析用到的变量有缺失值才去除该记录 (Excludes cases analysis by analysis)，或只要相关变量有缺失值，则在所有分析中均将该记录去除 (Excludes cases listwise)。默认为前者，以充分利用数据。

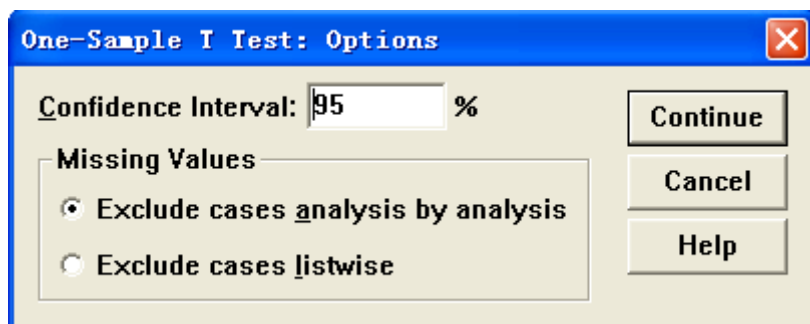


图 1.2b One-Samples T Test 的 Options 对话框

1.2.2 分析案例

比如要检验数据 111.sav 中 q9（消费）的总体均值是否等于 200。采用 One-Samples T Test 的简要操作步骤如下：

1. Analyze→Compare Means→One-Samples T Test
2. Test Variable (s) 框：选入 q9
3. Test Value 框：填入 200
4. 单击 OK

1.2.3 结果解释

One-Samples T Test 过程的输出也是比较简单的，由描述统计表和 t 检验表组成。上例的输出如下：

●One-Sample Statistics 分析表。所分析变量的基本情况描述，有样本量、均值、标准差和标准误。

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
被访问者最近一次参加促销活动的消费	312	114.03	18.158	3.887

●单样本 t 检验表，第一行注明了用于比较的已知总体均值为 200，下面从左到右依次为 t 值(t)、自由度(df)、P 值 (Sig. 2-tailed)、两均值的差值 (Mean Difference)、差值的 95%可信区间。由上表可知：t=-9.253，P=0.000<0.05。因此可以认为消费的总体均值不等于 200。

One-Sample Test

	Test Value = 200					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
被访问者最近一次参加促销活动的消费	-9.253	311	.000	-35.97	-43.12	-28.32

1.3 Independent-Samples T Test 过程

Independent-Samples T Test 过程用于进行两样本均值的比较，即常用的两样本 t 检验。该对话框的界面和上面的 One-Samples T Test 对话框非常相似。

1.3.1 界面说明

选择 Analyze→Compare Means→Independent-Samples T Test，进入对话框，见图 1.3a。其各部分解释如下：

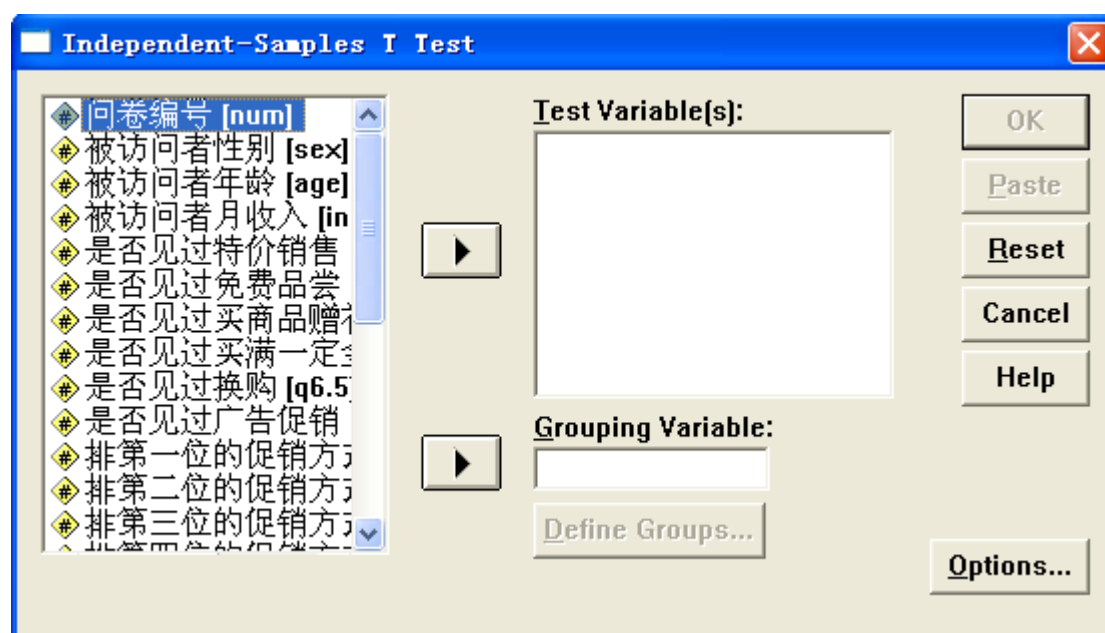


图 1.3a Independent-Samples T Test 对话框

- Test Variables 框：用于选入需要分析的变量。
- Grouping Variable 框：用于选入分组变量。注意选入变量后还要定义需比较的组别。
- Define Groups：单击后进入对话框（见图 1.3b），用于定义需要相互比较的两组的分组变量值。如果分组变量有 3 个取值（即有三组），而这个 t 检验是比较其中的某两组，这时就可以用 Define Groups 框来指定需比较的两组。当然，如果分组变量只有 2 个取值时，

仍然要再该框中进行定义，这也算是 SPSS 对话框存在的一个小缺陷吧。

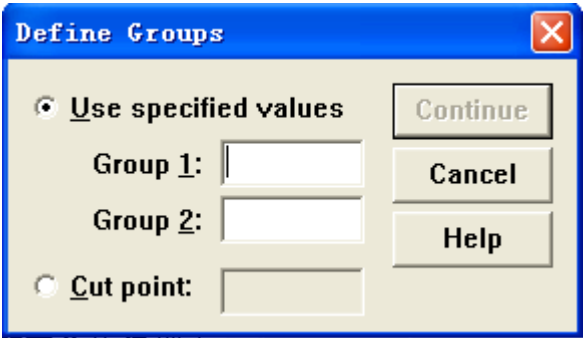


图 1.3b Independent-Samples T Test 的 Define Groups 对话框

●Options: 和 One-Samples T Test 对话框的 Options 完全相同，此处不再重复。

1.3.2 分析案例

要求检验数据 111.sav 中不同年龄组的消费 q9 是否相同。当然只能两个年龄组相比，如比较 25 岁以下与 25-35 岁两个组的消费均值是否相同。采用 Independent-Samples T Test 的简要操作步骤如下：

1. Analyze→Compare Means→Independent-Samples T Test
2. Test Variable (s) 框：选入 q9
3. Grouping Variable 框：选入 age
4. 单击 Define Groups: 在 Group1 框内输入 1, Group2 框内输入 2, 然后单击 Continue
5. 单击 OK

1.3.3 结果解释

用 Independent-Samples T Test 过程的结果输出如下：

●两组需检验变量的基本情况描述。

Group Statistics

	被访问者年龄	N	Mean	Std. Deviation	Std. Error Mean
被访问者最近一次参	25岁以下	53	110.12	11.441	9.121
加促销活动的消费	25-35岁	188	113.39	18.385	4.987

●Independent Samples Test 分析表。该结果分为两大部分：第一部分为 Levene's 方差齐性检验，用于判断两总体方差是否齐，这里的检验结果为 $F=0.251$, $P=0.113>0.05$ ，可见在本例中方差是齐的；第二部分则分别给出两组所在总体方差齐和方差不齐时的 t 检验结果，由于前面的方差齐性检验结果为方差齐，第二部分就应选用方差齐时的 t 检验结果，即上面一行列出的 $t=-0.212$, $df=239$, $P=0.793>0.05$ ，从而拒绝 H_0 ，认为这两个年龄组的消费没



什么不同。从上面的统计结果看，两个样本均值相差无几，也可认为两个组的消费无显著差异。最后面还附有一些其他指标，如两组均值的可信区间等，以对差异情况有更直观的了解。

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
被访问者最近一次参加促销活动的消费	Equal variances assumed	.256	.613	-.262	239	.793	-2.77	10.570	-23.594	18.052
	Equal variances not assumed			-.266	85.575	.791	-2.77	10.400	-23.448	17.906

1.4 Paired-Samples T Test 过程

该过程用于进行配对设计的样本差值均值与总体离差均值 0 比较的 t 检验，它和 One-Samples T Test 过程相重复的（等价于已知总体均值为 0 的情况），但 Paired-Samples T Test 过程使用的数据输入格式和前者不同，即通常所称的统计表格格式，因此仍然有存在的价值。

1.4.1 界面说明

选择 Analyze→Compare Means→Paired-Samples T Test，即可进入对话框，见图 1.4。整个界面上只有一个 Paired Variable 框需要介绍，它用于选入希望进行比较的一对或几对变量（注意这里的量词是对而不是个）。选入变量需要成对成对的选入，即按住 Ctrl 键，选中两个成对变量，再单击将其选入。如果只选中一个变量，则按钮为灰色，不可用。

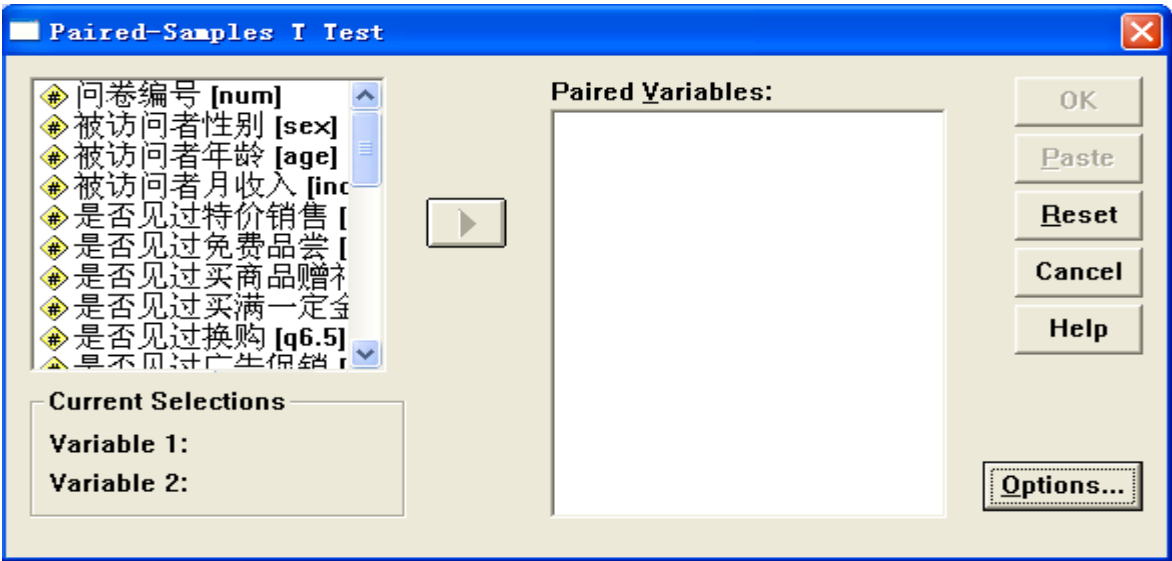


图 1.4 Paired-Samples T Test 对话框

1.4.2 分析实例

例 1.2 某单位研究饮食中缺乏维生素 E 与肝中维生素 A 含量的关系,将同种属的大白按性别相同,年龄、体重相近者配成对子,共 8 对,并将每对中的两头动物随机分到正常饲料组和维生素 E 缺乏组,过一定时期将大白鼠杀死,测得其肝中维生素 A 的含量,问不同饲料的大白鼠肝中维生素 A 含量有无差别?

大白鼠对号	正常饲料组	维生素 E 缺乏
1	3550	2450
2	2000	2400
3	3000	1800
4	3950	3200
5	3800	3250
1	3750	2700
7	3450	2500
8	3050	1750

为了说明问题,此处假设输入数据时就按照上表格式输入,其中正常饲料组变量名为 G1,维生素 E 缺乏组变量名为 G2。操作如下:

1. 同时选中 G1、G2: 选入 Paired Variables 框
2. 单击 OK

1.4.3 结果解释

以例 1.1 为例,其输出结果如下:

- 配对变量各自的统计描述,此处只有 1 对,故只有 Pair 1。

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	G1	3318.7500	8	632.4202	223.5943
	G2	2506.2500	8	555.1303	196.2682

- 此处进行配对变量间的相关性分析。等价于 Analyze→Correlate→Bivariate。

Paired Samples Correlations				
		N	Correlation	Sig.
Pair 1	G1 & G2	8	.584	.129

● 配对 t 检验表，给出最终的检验结果，由上表可见 $P=0.004$ ，故可认为两种饲料所得肝中维生素 A 含量有差别，即维生素 E 缺乏对大白鼠肝中维生素 A 含量有影响。

Paired Samples Test									
Paired Differences									
		95% Confidence Interval of the Difference							
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	G1 - G2	812.5000	546.2535	193.1298	355.8207	1269.1793	4.207	7	.004

上表的标题内容翻译如下：

		对子间的差异							
		差值均值	标准差	标准误	均值的 95%可信区间		t 值	自由 度	P 值（双侧）
					下限	上限			
第一 对	G1 - G2	812.5000	541.2535	193.1298	355.8207	1219.1793	4.207	7	.004

1.5 One-Way ANOVA 过程

One-Way ANOVA 过程用于进行两组及多组样本均值的比较，即成组设计的方差分析，如果做了相应选择，还可进行随后的两两比较，甚至于在各组间精确设定哪几组和哪几组进行比较，在本章的内容中，它是最为复杂的一个，但是有了前面的基础，拿下他应该不成问题。

1.5.1 界面说明

选择 Analyze→Compare Means→Paired-One-Way ANOVA，即可进入对话框，见图 1.5a。其各部分的解释如下：

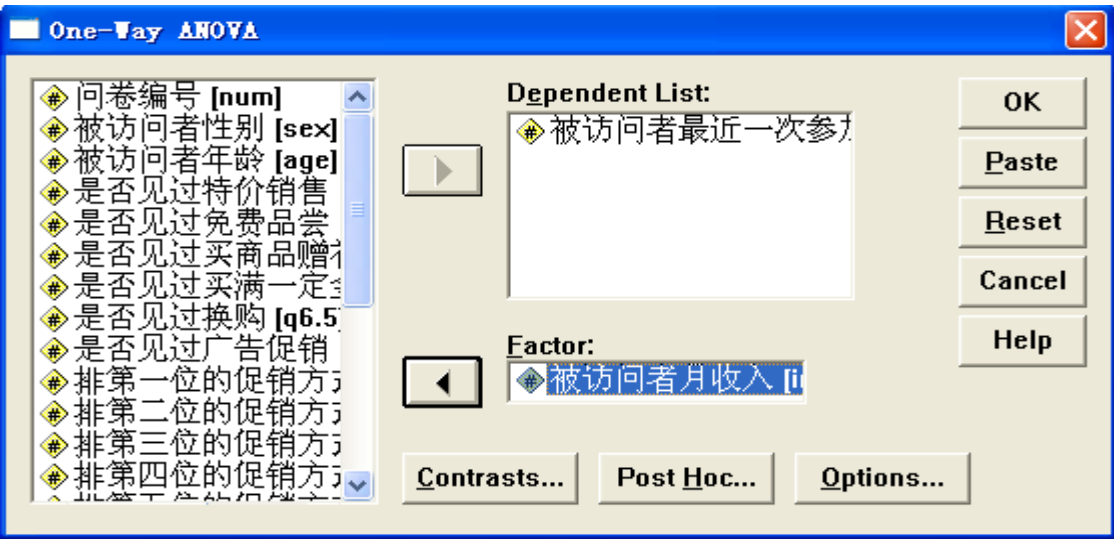


图 1.5a Paired-One-Way ANOVA 对话框

● **Dependent List 框**：选入需要分析的变量，可选入多个结果变量（因变量）。

● **Factor 框**：选入需要比较的分组因素，只能选入一个。

● **Contrast**：弹出 **Contrast** 对话框（见图 1.5b），用于对精细趋势检验和精确两两比较的选项进行定义，由于该对话框太专业，也较少用，这里只做简单介绍。

- **Polynomial 复选框**：定义是否在方差分析中进行趋势检验。
- **Degree 下拉列表**：和 Polynomial 复选框配合使用，可选则从线性趋势一直到最高五次方曲线来进行检验。
- **Coefficients 框**：定义精确两两比较的选项。这里按照分组变量升序给每组一个系数值，注意最终所有系数值相加应为 0。如果不为 0 仍可检验，只不过结果是错的。比如说在下面的例 1.2 中要对第一、三组进行单独比较，则在这里给三组分配系数为 1、0、-1，就会在结果中给出相应的检验内容。

● **Post Hoc**：弹出 **Post Hoc Multiple Comparisons** 对话框（见图 1.5c），用于选择进行各组间两两比较的方法，有：

- **Equar Variances Assumed 复选框组**：一组当各组方差齐时可用的两两比较方法，共有 14 种。这里不一一列出了，其中最常用的为 LSD、S-N-K、Tukey 等。
- **Equar Variances Not Assumed 复选框组**：一组当各组方差不齐时可用的两两比较方法，共有 4 种，其中以 Dunnetts's C 法较常用。
- **Significance Level 框**：定义两两比较时的显著性水平，默认为 0.05。

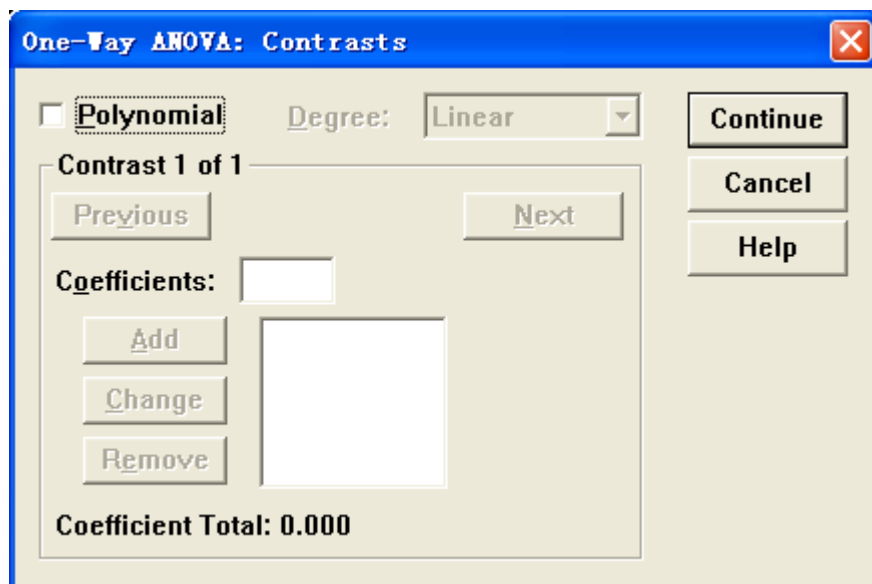


图 1.5b One-Way ANOVA 中的 Contrasts 对话框

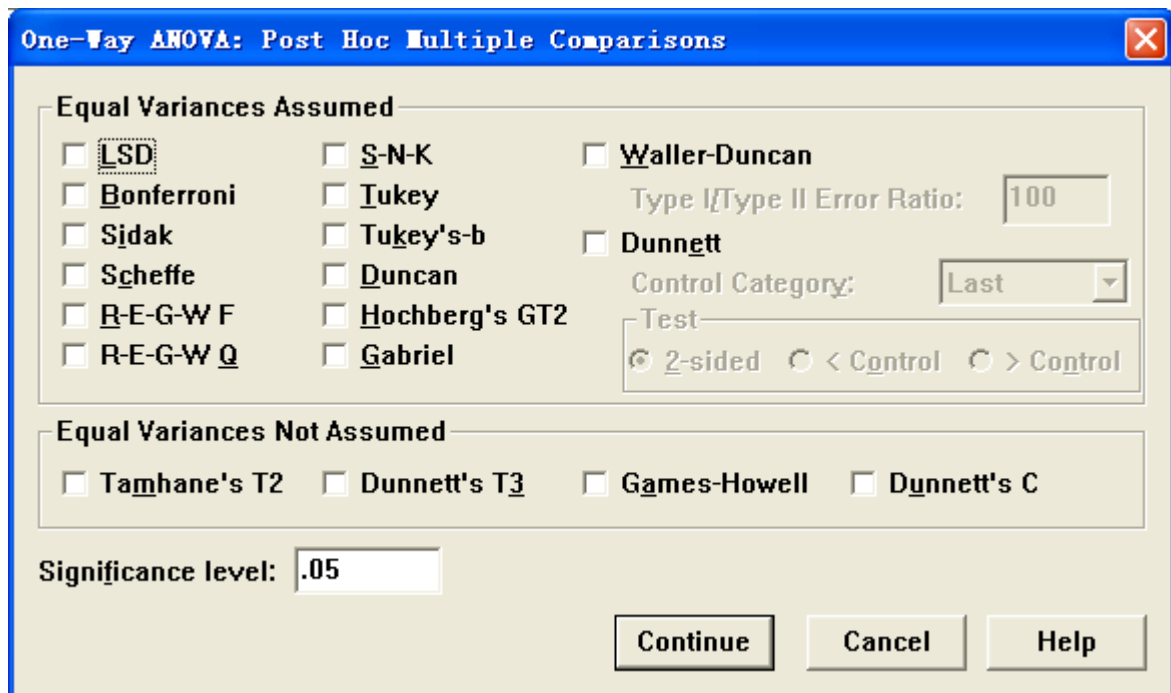


图 1.5c One-Way ANOVA 中的 Post Hoc Multiple Comparisons 对话框

●Options: 弹出 Options 对话框（见图 1.5d），用于定义相关的选项，有：

- Statistics 复选框组：选择一些附加的统计分析项目，有统计描述（Descriptive）和方差齐性检验（Homogeneity-of-variance）。
- Means plot 复选框：用各组均值做图，以直观的了解它们的差异。
- Missing Values 单选框组：定义分析中对缺失值的处理方法，可以是具体分析用到的变量有缺失值才去除该记录（Excludes cases analysis by analysis），或只要相关变量有缺失值，则在所有分析中均将该记录去除（Excludes cases listwise）。默认为前者，以充分利用数据。

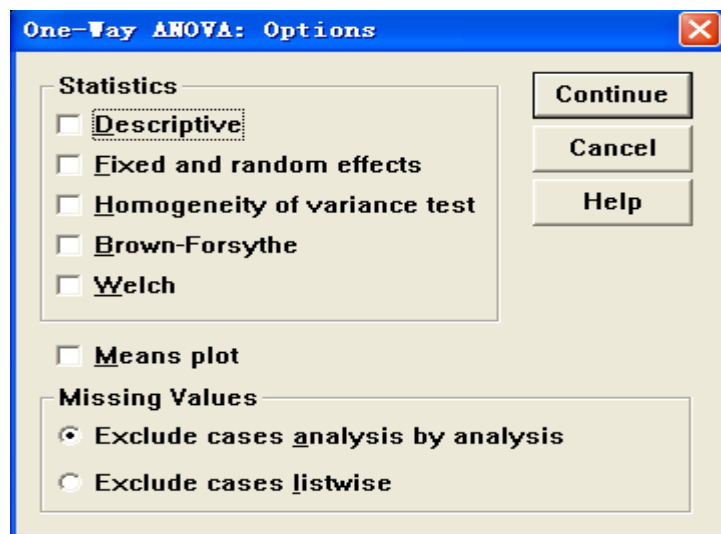


图 1.5c One-Way ANOVA 中的 Options 对话框

1.5.2 分析实例

例 1.3 利用 111.sav 文件中 q9（消费）、income（月收入）数据，研究四种收入群体的消费是否显著不同。

设 111.sav 数据文件已打开，分组变量为 income，因变量为 q9。此处先进行单因素方差分析，然后进行两两比较，这里选择 Tukey 法进行两两比较。操作如下：

1. Analyze→Compare Means→Paired-One-Way ANOVA
2. Dependent List 框：选入 q9
3. Factor 框：选入 income
4. 单击 Post Hoc：选中 Tukey 复选框，单击 Continue
5. 单击 OK

1.5.3 结果解释

上题的输出结果如下：

●一个典型的方差分析表。给出了单因素方差分析的结果，可见 $F=1.001$ ， $P=0.390>0.05$ 。因此可认为四组收入群体的最近一次参加促销活动的消费无显著差异。

ANOVA

被访问者最近一次参加促销活动的消费

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	14222.725	3	4740.908	1.001	.390
Within Groups	1451798.955	308	4713.133		
Total	1411021.179	311			

上表的标题内容翻译如下：

	离均差平方和 SS	自由度	均方 MS	F 值	P 值
组间变异	14222.725	3	4740.908	1.001	.390
组内变异	1451798.955	308	4713.133		
总变异	1411021.179	311			

●用 Tukey 法进行两两比较的结果。简单的说，在表格的纵向上有各配对组的均值差异、标准差、P 值及 95%的置信区间，表格的横向上被分成了若干个亚组。表中结果显示：不同亚组间的 P 值都大于 0.05，表明各组间两两比较均无有显著差异，可认为不同收入群体的消费几乎趋同。

Post Hoc Tests

Multiple Comparisons

Dependent Variable: 被访问者最近一次参加促销活动的消费

Tukey HSD

(I) 被访问者月收入	(J) 被访问者月收入	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1000元以下	1000-1500元	9.55	8.901	.701	-13.45	32.51
	1500-2000元	-5.55	11.418	.913	-35.17	24.07
	2000元以上	21.09	18.050	.147	-25.54	17.72
1000-1500元	1000元以下	-9.55	8.901	.701	-32.51	13.45
	1500-2000元	-15.11	12.181	.102	-41.57	11.31
	2000元以上	11.54	18.512	.925	-31.28	59.31
1500-2000元	1000元以下	5.55	11.418	.913	-24.07	35.17
	1000-1500元	15.11	12.181	.102	-11.31	41.57
	2000元以上	21.14	19.872	.538	-24.19	77.97
2000元以上	1000元以下	-21.09	18.050	.147	-17.72	25.54
	1000-1500元	-11.54	18.512	.925	-59.31	31.28
	1500-2000元	-21.14	19.872	.538	-77.97	24.19

相关分析 Correlate

在市场研究中经常要遇到分析两个或多个变量间关系的情况,有时是希望了解某个变量对另一个变量的影响强度,有时则是要了解变量间联系的密切程度,前者用下一章将要讲述的回归分析来实现,后者则需要用到本章所要讲述的相关分析实现。

SPSS 的相关分析功能被集中在 Statistics 菜单的 Correlate 子菜单中,他一般包括以下三个过程:

1. Bivariate 过程 此过程用于进行两个/多个变量间的参数/非参数相关分析,如果是多个变量,则给出两两相关的分析结果。这是 Correlate 子菜单中最为常用的一个过程,实际上对他的使用可能占到相关分析的 95%以上。下面的讲述也以该过程为主。
2. Partial 过程 如果需要进行相关分析的两个变量其取值均受到其他变量的影响,就可以利用偏相关分析对其他变量进行控制,输出控制其他变量影响后的相关系数,这种分析思想和协方差分析非常类似。Partial 过程就是专门进行偏相关分析的。
3. Distances 过程 调用此过程可对同一变量内部各观察单位间的数值或各个不同变量间进行距离相关分析,前者可用于检测观测值的接近程度,后者则常用于考察预测值对实际值的拟合优度。该过程在实际应用中用的非常少。

这里只介绍 Bivariate 过程。

2.1 Bivariate 过程

2.1.1 界面说明

选择 Analyze→Correlate→Bivariate，就可进入对话框，如图 2.1a 所示。其中各部分解释如下：

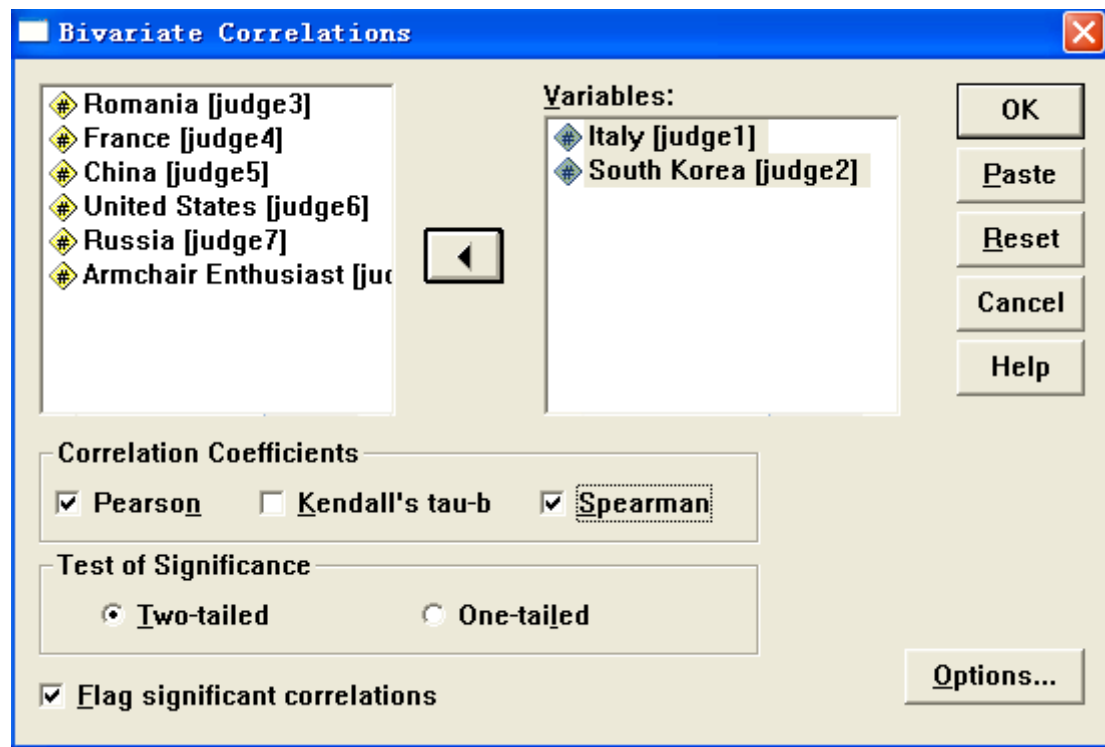


图 2.1a Bivariate 对话框

- Variables 框：用于选入需要进行相关分析的变量，至少需要选入两个。
- Correlation Coefficients 复选框组：用于选择需要计算的相关分析指标，有：
 - Pearson 复选框 选择进行积距相关分析，即最常用的参数相关分析
 - Kendall's tau-b 复选框 计算 Kendall's 等级相关系数
 - Spearman 复选框 计算 Spearman 相关系数，即最常用的非参数相关分析（秩相关）
- Test of Significance 单选框组：用于确定是进行相关系数的单侧（One-tailed）或双侧（Two-tailed）检验，一般选双侧检验。
- Flag significant correlations：用于确定是否在结果中用星号标记有统计学意义的相关系数，一般选中。此时 $P < 0.05$ 的系数值旁会标记一个星号， $P < 0.01$ 的则标记两个星号。
- Options：弹出 Options 对话框（见图 2.1b），选择需要计算的描述统计量和统计分析：

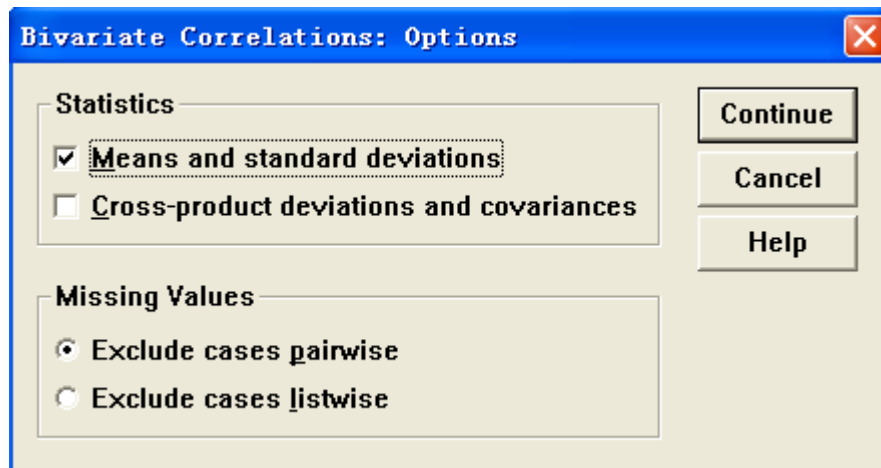


图 2.1b Bivariate Correlations 中的 Options 对话框

- **Statistics** 复选框组 可选的描述统计量。它们是：
 - Means and standard deviations 每个变量的均值和标准差
 - Cross-product deviations and covariances 各对变量的交叉积和以及协方差阵
- **Missing Values** 单选框组 定义分析中对缺失值的处理方法, 可以是具体分析用到的两个变量有缺失值才去除该记录 (Exclude cases pairwise), 或只要该记录中进行相关分析的变量有缺失值 (无论具体分析的两个变量是否缺失), 则在所有分析中均将该记录去除 (Excludes cases listwise)。默认为前者, 以充分利用数据。

2.1.2 分析实例

例 2.1 请计算 SPSS 自带的样本数据 judges.sav 中意大利 (judge1) 和韩国法官 (judge2) 得分的相关性。

由于 judge1 和 judge2 的数据分布不太好, 这里同时计算 Pearson 相关系数和 Spearman 相关系数。操作如下:

1. Variables 框: 选入 judge1、judge2
2. Pearson 复选框: 选中
3. Spearman 复选框: 选中
4. 单击 OK 钮

2.1.3 结果解释

例 2.1 的输出结果如下所示:

● **Pearson 相关系数系数表**。变量间两两的相关系数是用方阵的形式给出的。每一行和每一列的两个变量对应的格子中就是这两个变量相关分析结果结果, 共分为三列, 分别是相关系数、P 值和样本数。由于这里只分析了两个变量, 因此给出的是 2*2 的方阵。由上表可见 judge1、judge2 自身的相关系数均为 1 (of course), 而 judge1 和 judge2 的相关系数为 0.91, $P < 0.01$, 有非常显著的统计学意义。

如果需要得到具体的 P 值。请进入表格的编辑模式, 双击 P 值所在的单元格, 就可以看

到精确的 P 值大小。

Correlations

Correlations			
		Italy	South Korea
Italy	Pearson Correlation	1.000	.910**
	Sig. (2-tailed)	.	.000
	N	300	300
South Korea	Pearson Correlation	.910**	1.000
	Sig. (2-tailed)	.000	.
	N	300	300

**:. Correlation is significant at the 0.01 level (2-tailed).

上表的标题内容翻译如下：

		Italy	South Korea
Italy	Pearson 积距相关系数	1.000	.910
	P 值（双侧）		.000
	样本数	300	300
South Korea	Pearson 积距相关系数	.910	1.000
	P 值（双侧）	.000	
	样本数	300	300

●Spearman 相关系数系数表。此处的表格内容和上面 Pearson 相关系数的结果非常相似，只是表格左侧注明为 Spearman 等级相关。可见 judge1 和 judge2 的等级相关系数为 0.92，P<0.001，有非常显著的统计学意义。

Nonparametric Correlations

Correlations				
			Italy	South Korea
Spearman's rho	Italy	Correlation Coefficient	1.000	.920**
		Sig. (2-tailed)	.	.000
		N	300	300
	South Korea	Correlation Coefficient	.920**	1.000
		Sig. (2-tailed)	.000	.
		N	300	300

**:. Correlation is significant at the .01 level (2-tailed).

回归分析：线性回归与曲线拟合 Regression

回归分析是处理两个及两个以上变量间线性依存关系的统计方法。在现实问题研究中，这类方法用的比较多。如产品销售与广告费用有关系，人头发中某种金属元素的含量与血液中该元素的含量有关系，人的体表面积与身高、体重有关系；等等。回归分析就是用于说明这种依存变化的数学关系。

3.1 Linear 过程

3.1.1 简单操作入门

调用此过程可完成二元或多元的线性回归分析。在多元线性回归分析中，用户还可根据需要，选用不同筛选自变量的方法（如：逐步法、向前法、向后法，等）。

例 3.1 请分析在数据集 444. sav 中变量 X1（化肥施肥量）对变量 Y（粮食产量）的大小有无影响？或者 X1、X2、X3、X4、X5 变化对 Y 有无影响？

显然，这里所有变量都是连续性变量，用单因素方差分析或多因素方差分析不太现实，应采用回归分析来解决。前一个问题是双变量回归问题，而后面是多元回归问题。其实，回归分析和方差分析都可以被归入广义线性模型中，因此他们在模型的定义、计算方法等许多方面都非常近似，下面大家很快就会看到。

3.1.1.1 界面详解

在菜单中选择 Regression→linear，系统弹出线性回归对话框，如图 3.1a 所示：

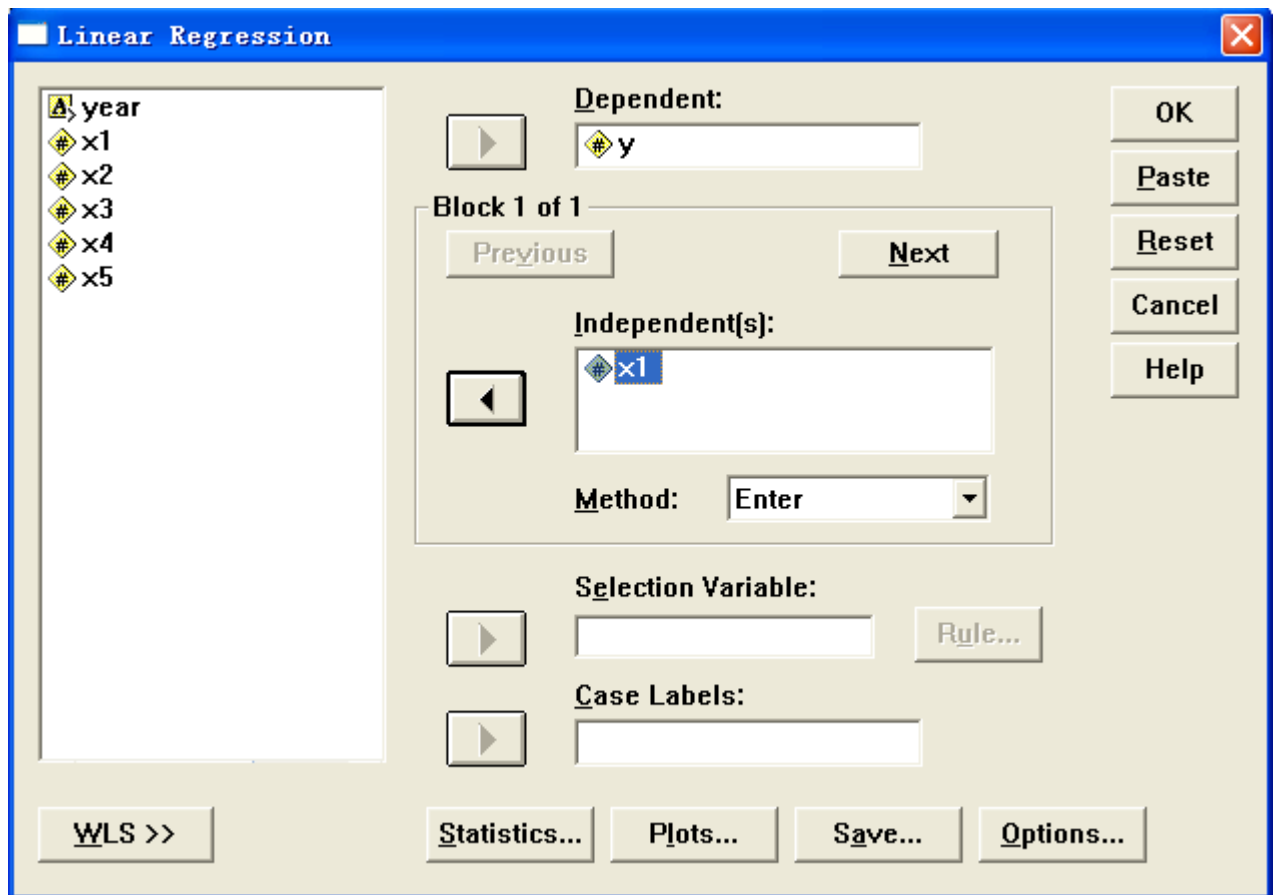


图 3.1a Linear Regression 对话框

除了大家熟悉的内容以外，里面还出现了一些特色项：

●Dependent 框：用于选入回归分析的因变量。

●Block 按钮组：由 Previous 和 Next 两个按钮组成，用于将下面 Independent 框中选入的自变量分组。由于多元回归分析中自变量的选入方式有前进、后退、逐步等方法，如果不同的自变量选入的方法不同，则用该按钮组将自变量分组选入即可。下面的例子会讲解其用法。

●Independent 框：用于选入回归分析的自变量。双变量回归时，选入一个变量即可；若多元回归时，就需要选入多个变量。由于是线性回归分析，所以要求所有选入的变量必须与 Dependent 框中的变量是线性相关，也就是回归方程必须是线性的。

●Method 下拉列表：用于选择对自变量的选入方法，有 Enter（强行进入法）、Stepwise（逐步法）、Remove（强制剔除法）、Backward（向后法）、Forward（向前法）五种。该选项对当前 Independent 框中的所有变量均有效。

●Selection Variable 框：选入一个筛选变量，并利用右侧的 Rules 钮建立一个选择条件，这样，只有满足该条件的记录才会进入回归分析。

●Case Labels 框：选择一个变量，它的取值将作为每条记录的标签。最典型的情况是使用记录 ID 号的变量。

●WLS>>：可利用该按钮进行权重最小二乘法的回归分析。单击该按钮会扩展当前对话框，出现 WLS Weight 框，在该框内选入权重变量即可。

●Statistics：弹出 Statistics 对话框（见图 3.1b），用于选择所需要的描述统计量。有如下选项：

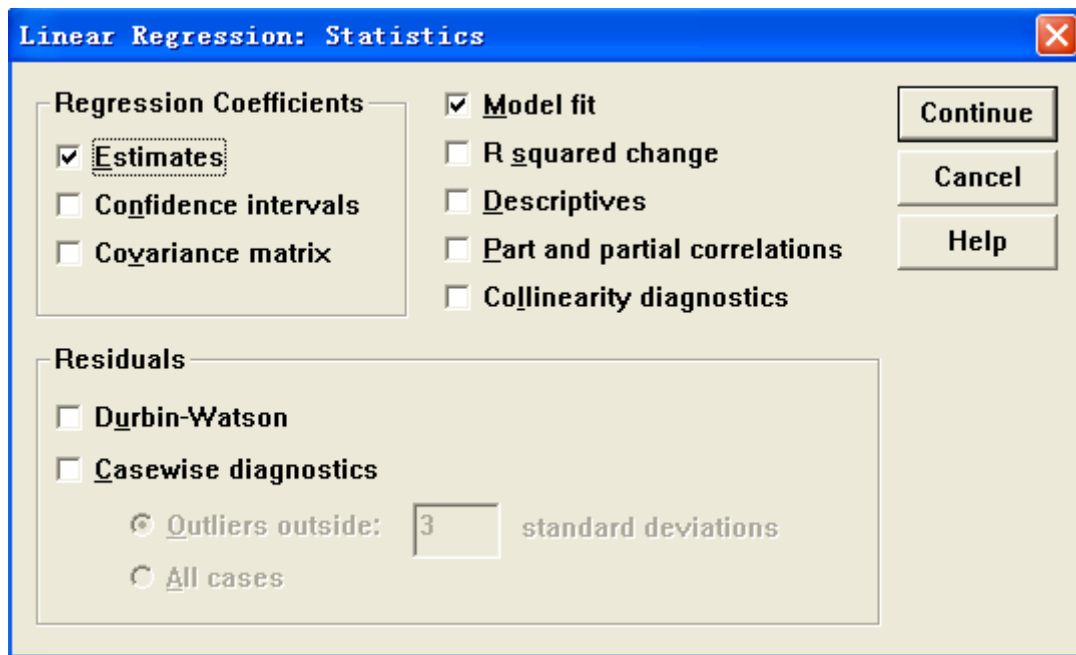


图 3.1 不 Linear Regression 中的 Statistics 对话框

- Regression Coefficients 复选框组：定义回归系数的输出情况，选中 Estimates 可输出回归系数 B 及其标准误，t 值和 p 值，还有标准化的回归系数 beta；选中 Confidence intervals 则输出每个回归系数的 95% 可信区间；选中 covariance matrix 则会输出各个自变量的相关矩阵和方差、协方差矩阵。以上选项默认只选中 Estimates。
- Residuals 复选框组：用于选择输出残差诊断的信息，可选的有 Durbin-Watson 残差序列相关性检验、超出规定的 n 倍标准误的残差列表。
- Model fit 复选框：模型拟合过程中进入、退出的变量的列表，以及一些有关拟合优度的检验：，R，R² 和调整的 R²，标准误及方差分析表。
- R squared change 复选框：显示模型拟合过程中 R²、F 值和 p 值的改变情况。
- Descriptives 复选框：提供一些变量描述，如有效例数、均值、标准差等，同时还给出一个自变量间的相关矩阵。
- Part and partial correlations 复选框：显示自变量间的相关、部分相关和偏相关系数。
- Collinearity diagnostics 复选框：给出一些用于共线性诊断的统计量，如特征根 (Eigenvalues)、方差膨胀因子 (VIF) 等。

以上各项在默认情况下只有 Estimates 和 Model fit 复选框被选中。

● Plot：弹出 Plot 对话框（见图 3.1c），用于选择需要绘制的回归分析诊断或预测图。可绘制的有标准化残差的直方图和正态分布图，因变量、预测值和各自变量残差间两两的散点图等。

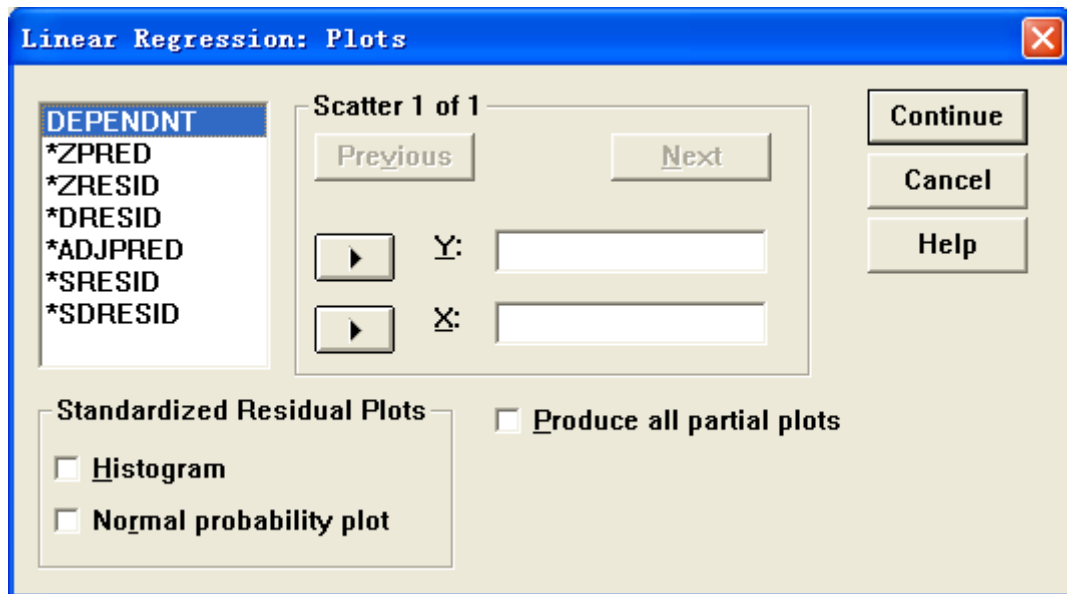


图 3.1c Linear Regression 中的 Plots 对话框

●Save: 许多时候需要将回归分析的结果存储起来，然后用得到的残差、预测值等做进一步的分析，Save 钮就是用来存储中间结果的。可以存储的有：预测值系列、残差系列、距离（Distances）系列、预测值可信区间系列、波动统计量系列，见图 3.1d。下方的按钮用于选择将这些新变量存储到一个新的 SPSS 数据文件或 XML 中。

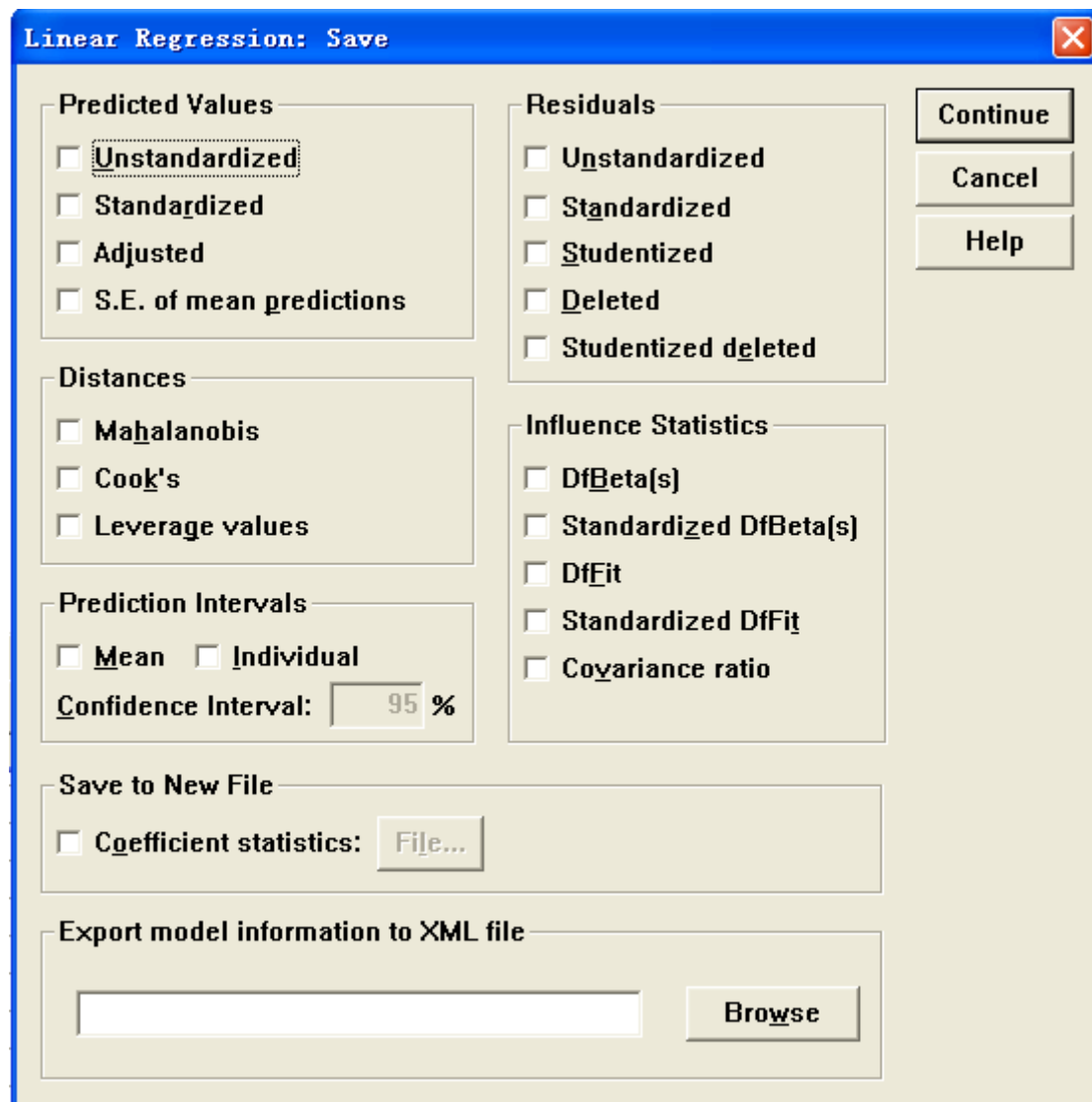


图 3.1d Linear Regression 中的 Save 对话框

●Options: 设置回归分析的一些选项，见图 3.1e:

- Stepping Method Criteria 单选钮组: 设置纳入和排除标准，可按 P 值或 F 值来设置。
- Include constant in equation 复选框: 用于决定是否在模型中包括常数项，默认选中。
- Missing Values 单选钮组: 用于选择对缺失值的处理方式，可以是不分析任一选入的变量有缺失值的记录 (Exclude cases listwise) 而无论该缺失变量最终是否进入模型; 不分析具体进入某变量时有缺失值的记录 (Exclude cases pairwise); 将缺失值用该变量的均值代替 (Replace with mean)。

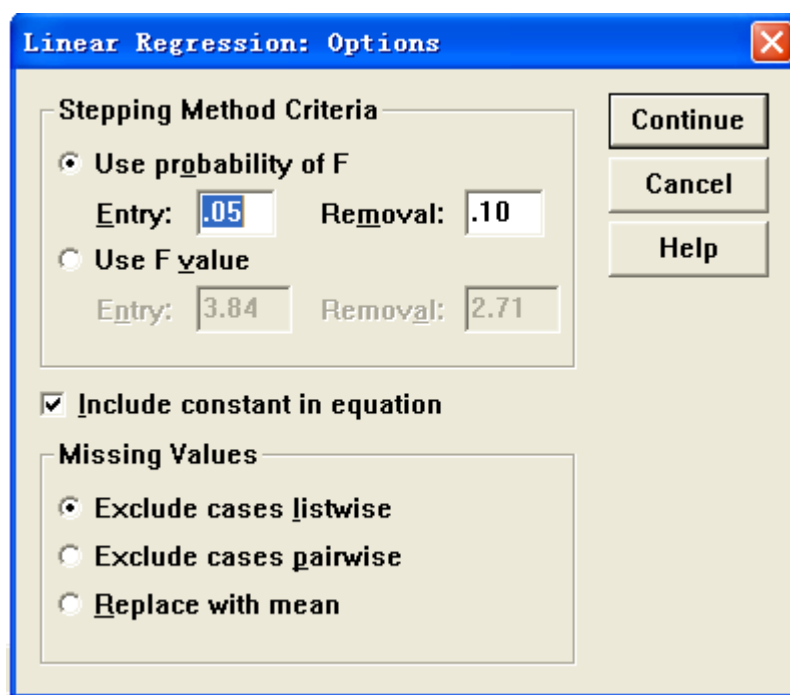


图 3.1e Linear Regression 中的 Options 对话框

3.1.1.2 输出结果解释

根据题目的要求，第一个问题只需在 Dependent 框中选入 Y，Independent 框中选入 X1 即可，其他的选项一律不管，单击 OK 就行了。而第二个问题是多元回归，除了在 Independent 框中选入所有 X 变量外（注意 Block 顺序），最好还要改 Method 框中的 Enter 为 Forward 或者 Backward，以检验多重共线性问题。此外，还可在 Statistics 中选 Durbin-Watson、Collinearity diagnostics 等。

下面是第一个问题的分析结果。第二个问题就由大家参照后面的实例去练习一下。

●这里的表格是拟合过程中变量进入/退出模型的情况记录，由于只引入了一个自变量，所以只出现了一个模型 1（在多元回归中就会依次出现多个回归模型），该模型中 X1 为进入的变量，没有移出的变量，具体的进入/退出方法为 enter。

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	X1(a)	.	Enter

a All requested variables entered.

b Dependent Variable: Y

●拟合模型的情况简报，显示在模型 1 中相关系数 R 为 0.994，而决定系数 R² 为 0.892，校正的决定系数为 0.885，说明模型的拟合度较高。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.944(a)	.892	.885	1493.98371

a Predictors: (Constant), X1

●这是所用模型的检验结果，可以看到这就是一个标准的方差分析表！有兴趣的话，可以自己用方差分析模型做一下，就会发现出了最左侧的一列名字不太一样外，其他的各个参数值都是相同的。从上表可见所用的回归模型 F 值为 132.017，P 值为 0.000，因此用的这个回归模型是有统计学意义的，可以继续看下面系数分别检验的结果。由于这里所用的回归模型只有一个自变量，因此模型的检验就等价与系数的检验，在多元回归中这两者是不同的。

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	294770998.380	1	294770998.380	132.017	.000(a)
	Residual	35711799.398	11	2231987.412		
	Total	330482797.778	17			

a Predictors: (Constant), X1

b Dependent Variable: Y

●包括常数项在内的所有系数的检验结果。用的是 t 检验，同时还会给出标化/未标化系数。可见常数项和 X1 都是有统计学意义的。

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	30817.311	1201.314		25.587	.000
	X1	4.571	.398	.944	11.492	.000

a Dependent Variable: Y

上表的内容如果翻译成中文则如下所示：

		未标准化系数		标准化系数		
模型		系数 b	系数标准误	系数 β	t 值	P 值
1	常数	30817.311	1201.314		25.587	.000
	X1	4.571	.398	.944	11.492	.000

3.1.2 复杂实例操作

3.1.2.1 分析实例

例 3.2 请分析在数据集 plastic.sav 中变量 extrusn、additive、gloss 和 opacity 对变量 tear_res 的大小有无影响？已知 extrusn 对 tear_res 的大小有影响。

显然，这里是一个多元回归，由于除了 extrusn 确有影响以外，不知道另三个变量有无影响，因此这里将 extrusn 放在第一个 block，进入方法为 enter（有把握 extrusn 一定有统计学意义）；另三个变量放在第二个 block，进入方法为 stepwise（让软件自动选择判断），操作如下：

- 1. Analyze→Regression→Linear
- 2. Dependent 框：选入 tear_res
- 3. Independent 框：选入 extrusn；单击 next 钮
- 4. Independent 框：选入 additive、gloss 和 opacity；Method 列表框：选择 stepwise
- 5. 单击 OK 钮

3.1.2.2 结果解释

最终的结果如下：
●表格依次列出了模型的筛选过程，模型 1 用进入法引入了 extrusn，然后模型 2 用 stepwise 法引入了 additive，另两个变量因没有达到进入标准，最终没有进入。

Regression

Variables Entered/Removed ^b			
Model	Variables Entered	Variables Removed	Method
1	Extrusion ^a	.	Enter
2	Additive Amount	.	Stepwise (Criteria: Probability-of-F-to-enter ≤ .050, Probability-of-F-to-remove ≥ .100).

a. All requested variables entered.
b. Dependent Variable: Tear Resistance

上面的表格翻译出来如下：

模型	进入的变量	移出的变量	变量筛选方法
1	extrusn		进入法
2	additive		stepwise 法（标准：进入概率小于 0.05，移出概率大于 0.1）

●两个模型变异系数的改变情况，从调整的 R2 可见，从上到下随着新变量的引入，模型可解释的变异占总变异的比例越来越大。

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.639 ^a	.408	.375	.375
2	.766 ^b	.586	.538	.322

a. Predictors: (Constant), Extrusion

b. Predictors: (Constant), Extrusion, Additive Amount

- 所用两个模型的检验结果，用的方法是方差分析，可见二个模型都有统计学意义。

ANOVA^c

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.740	1	1.740	12.408	.002 ^a
	Residual	2.525	18	.140		
	Total	4.266	19			
2	Regression	2.501	2	1.250	12.048	.001 ^b
	Residual	1.765	17	.104		
	Total	4.266	19			

a. Predictors: (Constant), Extrusion

b. Predictors: (Constant), Extrusion, Additive Amount

c. Dependent Variable: Tear Resistance

- 三个模型中各个系数的检验结果，用的是 t 检验，可见在模型 2 中所有的系数都有统计学意义。

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.900	.265		22.278	.000
	Extrusion	.590	.167	.639	3.522	.002
2	(Constant)	5.315	.314		16.926	.000
	Extrusion	.590	.144	.639	4.095	.001
	Additive Amount	.390	.144	.422	2.707	.015

a. Dependent Variable: Tear Resistance

上表的内容翻译如下：

		未标化的系数		标化的系数		
模型		B	标准误	Beta	t 值	P 值
1	(常数)	5.900	.215		22.278	.000
	extrusion	.590	.117	.139	3.522	.000
2	(常数)	5.315	.314		11.921	.000
	extrusion	.590	.144	.139	4.905	.000
	additive	.390	.144	.422	2.707	.000

●这是新出现的一个表格，反映的是没有进入模型的各个变量的检验结果，可见在模型 1 中，未引入模型的候选变量 additive 还有统计学意义，可能需要引入，而模型 2 中没有引入的两个变量其 P 值均大于 0.05，无需再进行分析了。

Excluded Variables ^c						
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	Gloss	.207 ^a	.986	.338	.233	.744
	Opacity	-.062 ^a	-.332	.744	-.080	.994
	Additive Amount	.422 ^a	2.707	.015	.549	1.000
2	Gloss	.013 ^b	.062	.952	.015	.624
	Opacity	-.183 ^b	-1.142	.270	-.274	.928

a. Predictors in the Model: (Constant), Extrusion

b. Predictors in the Model: (Constant), Extrusion, Additive Amount

c. Dependent Variable: Tear Resistance

3.2 Curve Estimation 过程

Curve Estimation 过程可以用与拟合各种各样的曲线，原则上只要两个变量间存在某种可以被它所描述的数量关系，就可以用该过程来分析。但这里要指出，由于曲线拟合非常的复杂，而该模块的功能十分有限，因此最好采用将曲线相关关系通过变量变换的方式转化为直线回归的形式来分析，或者采用其他专用的模块分析。

3.2.1 界面详解

Curve Estimation 过程中有特色的对话框界面内容，如图 3.2a 所示：

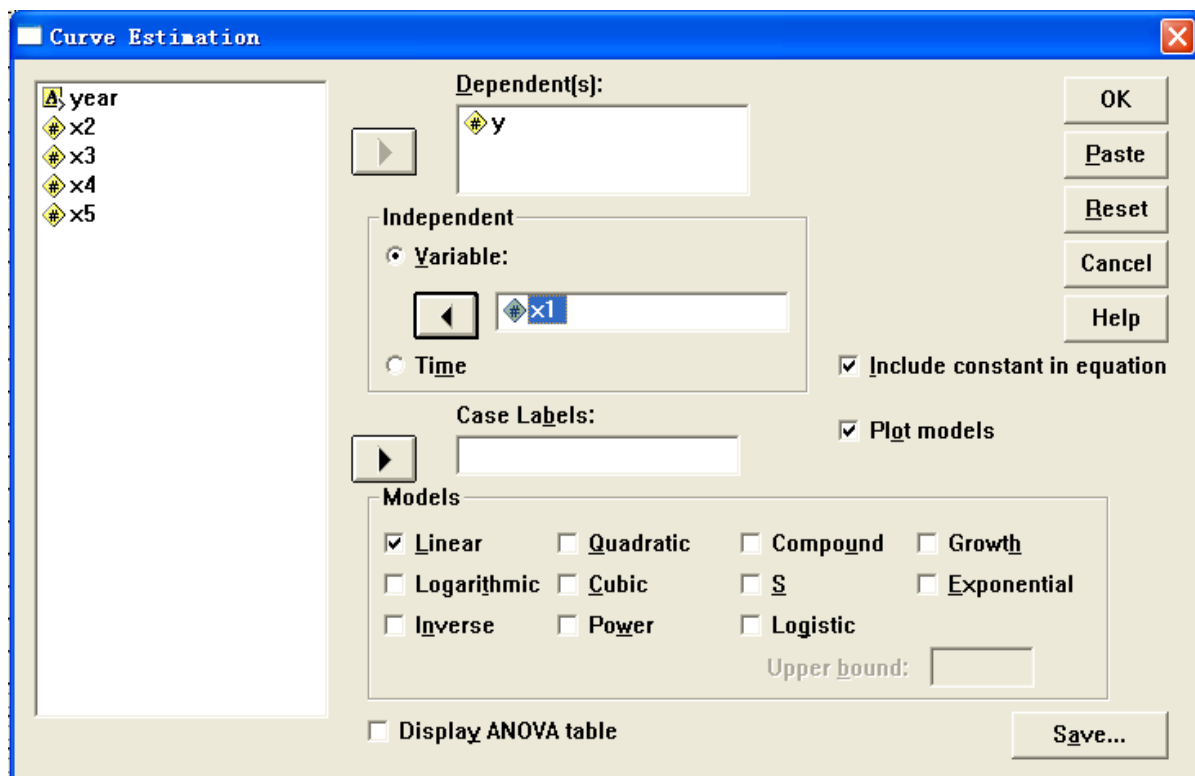


图 3.2a Curve Estimation 对话框

下面分别解释一下它们的具体功能：

●Dependent 框：用于选入曲线拟和中的因变量，可选入多个，如果这样，则对各个因变量分别拟合模型。

●Independent 单选框组：用于选入曲线拟和中的自变量，有两种选择，可以选入普通的自变量，也可以选择时间作为自变量，如果这样做，则所用的数据应为时间序列数据格式。

●Models 复选框组：是该对话框的重点，用于选择所用的曲线模型，可用的有：

- Linear：拟合直线方程，实际上与 Linear 过程的二元直线回归相同；
- Quadratic：拟合二次方程 $Y = b_0 + b_1X + b_2X^2$ ；
- Compound：拟合复合曲线模型 $Y = b_0 \times b_1X$ ；
- Growth：拟合等比级数曲线模型 $Y = e(b_0 + b_1X)$ ；
- Logarithmic：拟合对数方程 $Y = b_0 + b_1 \ln X$ ；
- Cubic：拟合三次方程 $Y = b_0 + b_1X + b_2X^2 + b_3X^3$ ；
- S：拟合 S 形曲线 $Y = e(b_0 + b_1/X)$ ；
- Exponential：拟合指数方程 $Y = b_0 e^{b_1X}$ ；
- Inverse：数据按 $Y = b_0 + b_1/X$ 进行变换；
- Power：拟合乘幂曲线模型 $Y = b_0X^{b_1}$ ；
- Logistic：拟合 Logistic 曲线模型 $Y = 1 / (1/u + b_0 \times b_1X)$ ，如选择该线型则要求输入上界。

上面的几种线型和其他的模块有重复，如 Logistic、Liner 等，由于本模块的功能有限，在重复的情况下建议用其它专用模块来分析。

- Include constant in equation 复选框：确定是否在方程中包含常数项。
- Plot models 复选框：要求对模型做图，包括原始数值的连线图和拟合模型的曲线图。
- Save：弹出 SAVE 对话框（见图 3.2b），用于定义想要存储的中间结果，如预测值、

预测值可信区间、残差等。

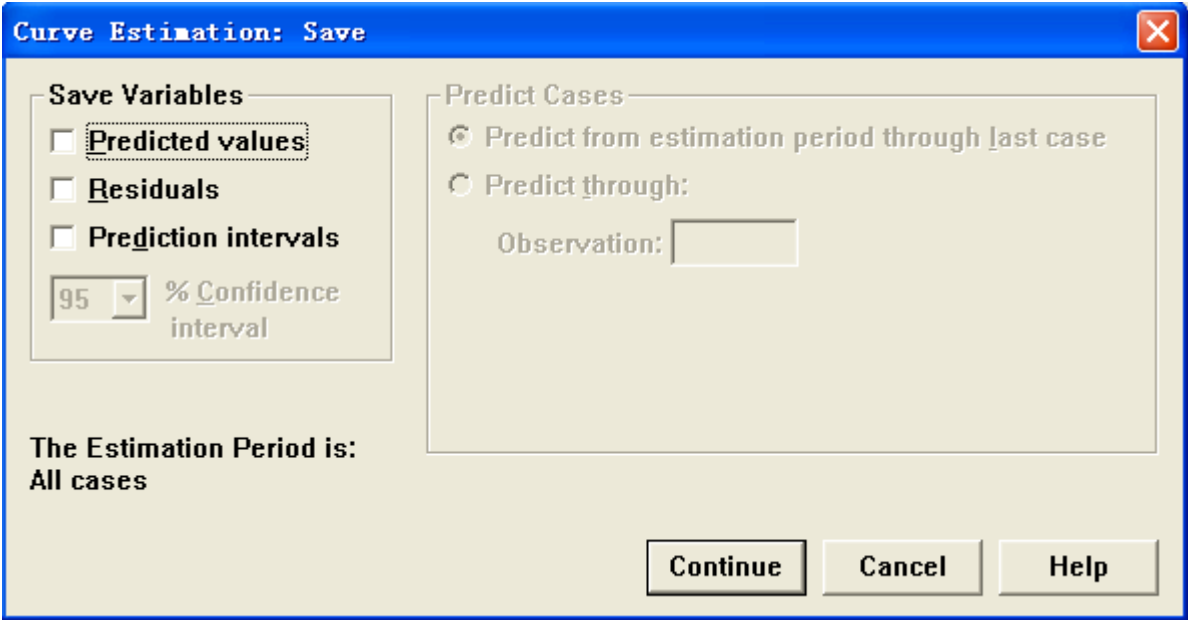


图 3.2b Curve Estimation 中的 Save 对话框

●Display ANOVA table 复选框：要求显示模型检验的方差分析表。

附：实验项目 3：均值比较与回归分析

实验项目	能应用 SPSS 软件进行：单个均值假设检验、均值比较、相关分析、回归分析
实验日期	
实验环境	SPSS for WINDOWS
实验内容	依据上个实验的数据文件或选择 SPSS 数据库中的文件，进行： 1、单个均值假设检验分析 2、均值比较分析 3、相关分析 4、回归分析
实验步骤	根据实验自己认真填写.
实 验 结 论 (或 实 验 体 会)	1. 写出求解问题的主要结果。 2. 谈谈实验体会。
实验批改	