知乎

SPSS操作：简单线性回归（史上最详尽的手把手教程）

医小咖
网站、微信公众号：医咖会，临床研究的传播者。

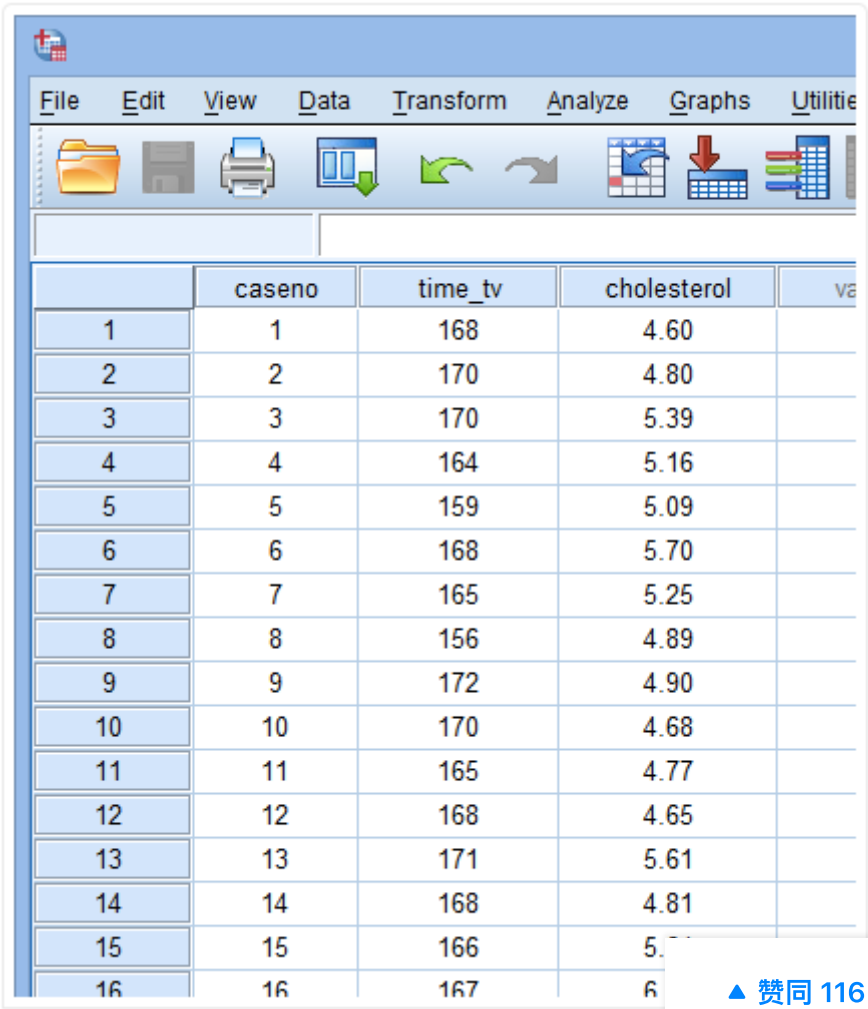
116 人赞同了该文章

1、问题与数据

研究表明，运动有助于预防心脏病。一般来说，运动越多，心脏病的患病风险越小。其原因之一在于，运动可以降低血胆固醇浓度。近期研究显示，一项久坐的生活指标—看电视时间，可能是罹患心脏病的预测因素。即看电视时间越长，心脏病的患病风险越大。

研究者拟在45-65岁健康男性人群中分析胆固醇浓度与看电视时间的关系。他们猜测可能存在正向相关，即看电视时间越长，胆固醇浓度越高。同时，他们也希望预测胆固醇浓度，并计算看电视时间对胆固醇浓度的解释能力。

研究者收集了受试者每天看电视时间（time_tv）和胆固醇浓度（cholesterol）等变量信息，部分数据如下：



	caseno	time_tv	cholesterol	va
1	1	168	4.60	
2	2	170	4.80	
3	3	170	5.39	
4	4	164	5.16	
5	5	159	5.09	
6	6	168	5.70	
7	7	165	5.25	
8	8	156	4.89	
9	9	172	4.90	
10	10	170	4.68	
11	11	165	4.77	
12	12	168	4.65	
13	13	171	5.61	
14	14	168	4.81	
15	15	166	5.	
16	16	167	6	

研究者想判断两个变量之间的关系，同时用其中一个变量（看电视时间）预测另一个变量（胆固醇浓度），并计算其中一个变量（看电视时间）对另一个变量（胆固醇浓度）变异的解释程度。针对这种情况，我们可以使用简单线性回归分析，但需要先满足7项假设：

假设1：因变量是连续变量

假设2：自变量可以被定义为连续变量

假设3：因变量和自变量之间存在线性关系

假设4：具有相互独立的观测值

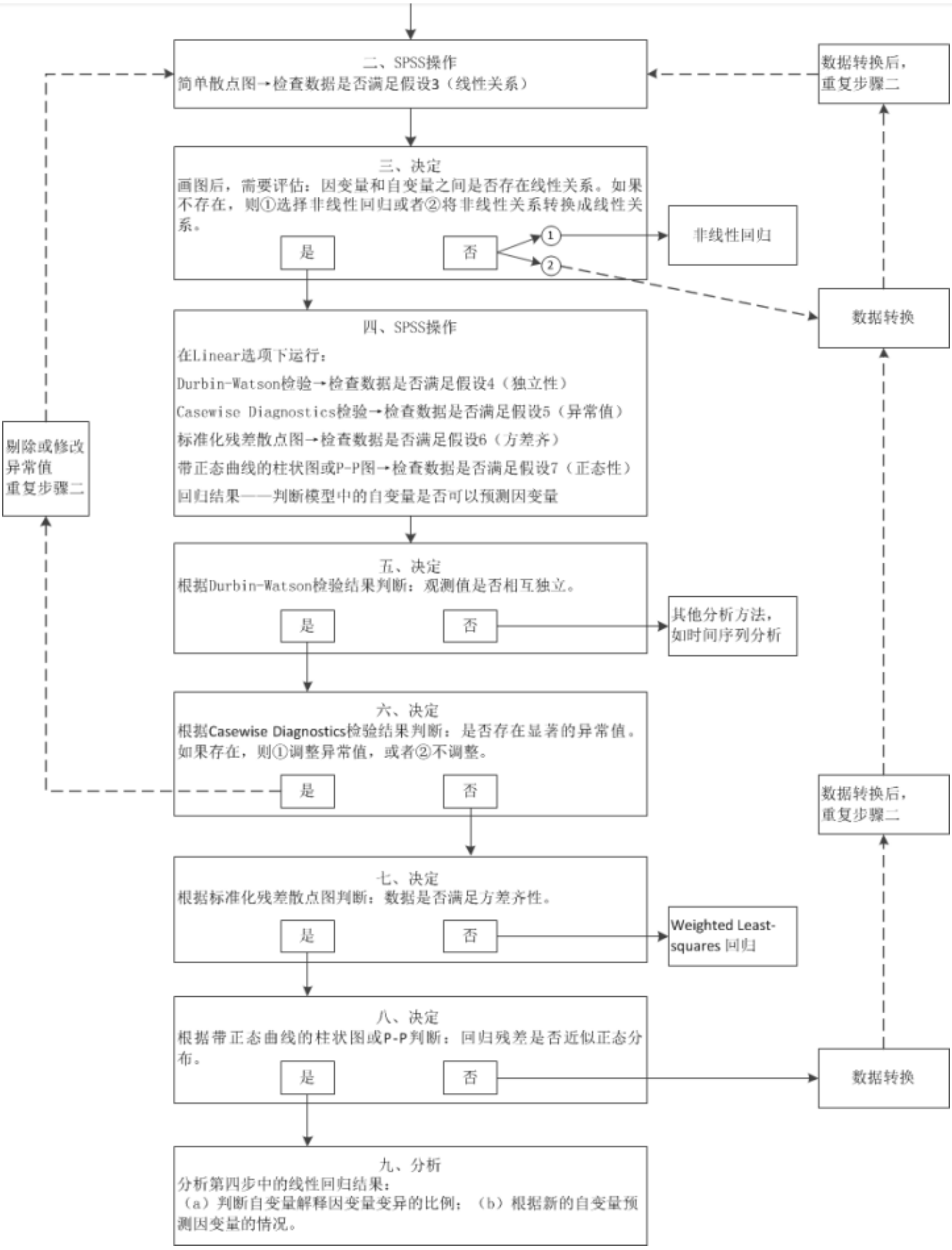
假设5：不存在显著的异常值

假设6：等方差性

假设7：回归残差近似正态分布

那么，进行简单线性回归分析时，如何考虑和处理这7项假设呢？

3、思维导图



4、对假设的判断

因变量是连续变量，自变量可以被定义为连续变量。

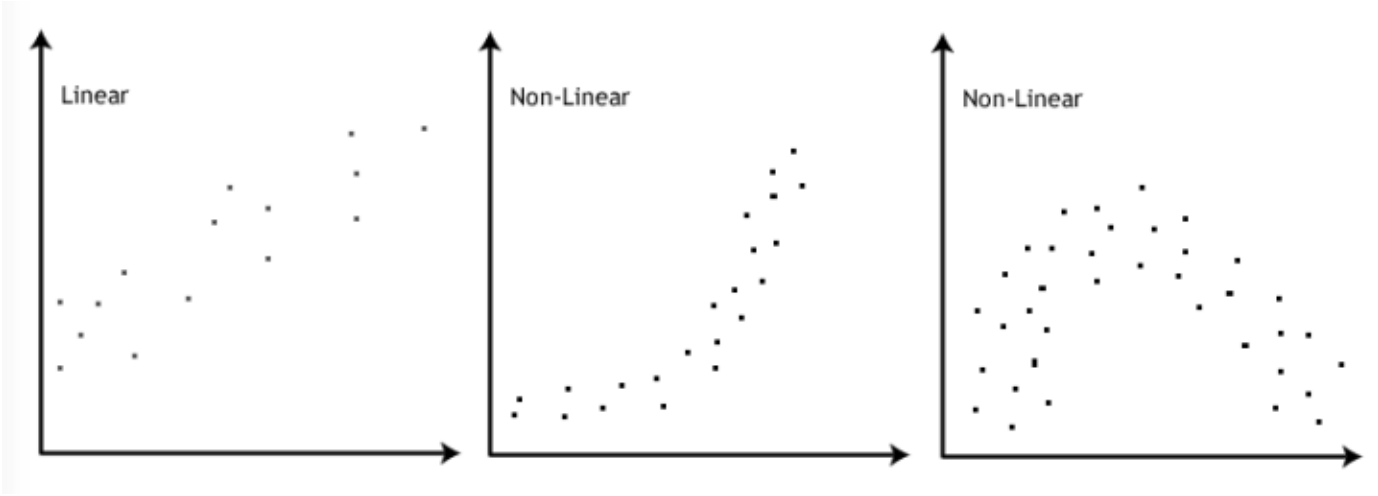
举例来说，我们平时测量的反应时间（小时）、智力水平（IQ分数）、考试成绩（0到100分）以及体重（千克）都是连续变量。在线性回归中，因变量（dependent variable）一般是指研究的成果、目标或者标准值；自变量（independent variable）一般被看作预测、解释或者回归变量。

假设1和假设2与研究设计有关，需要根据实际情况判断。

4.2 假设3

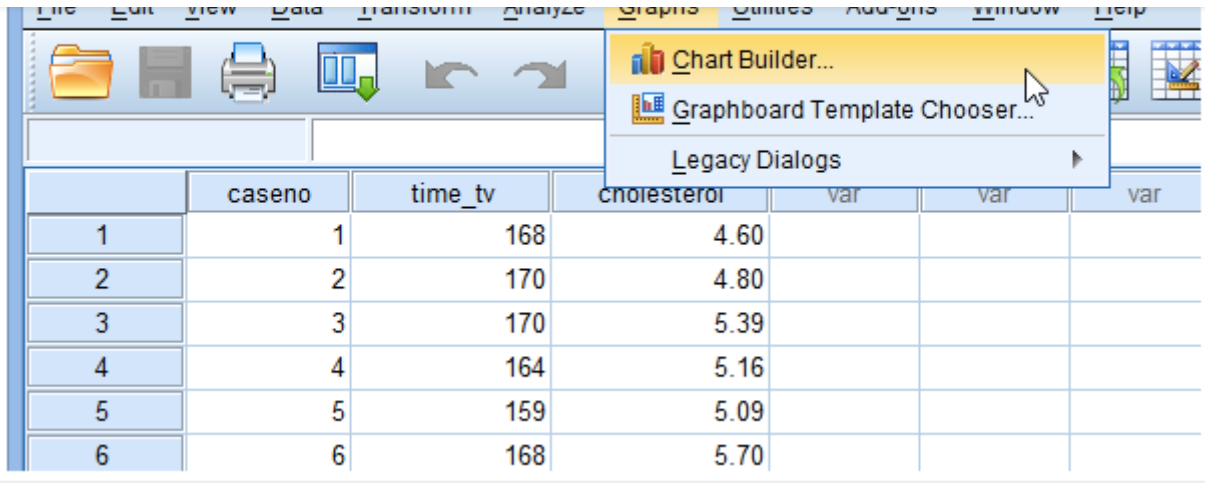
简单线性回归要求自变量和因变量之间存在线性关系，如要求看电视时间（time_tv）和胆固醇浓度（cholesterol）存在线性关系。

判断变量之间是否存在线性关系的方法有很多，我们主要向大家介绍散点图法，即通过因变量和自变量的散点图进行直观地判断。如果散点趋向于构成一条直线，那么因变量和自变量之间存在线性关系；如果构成曲线，就不存在线性关系，举例如下：



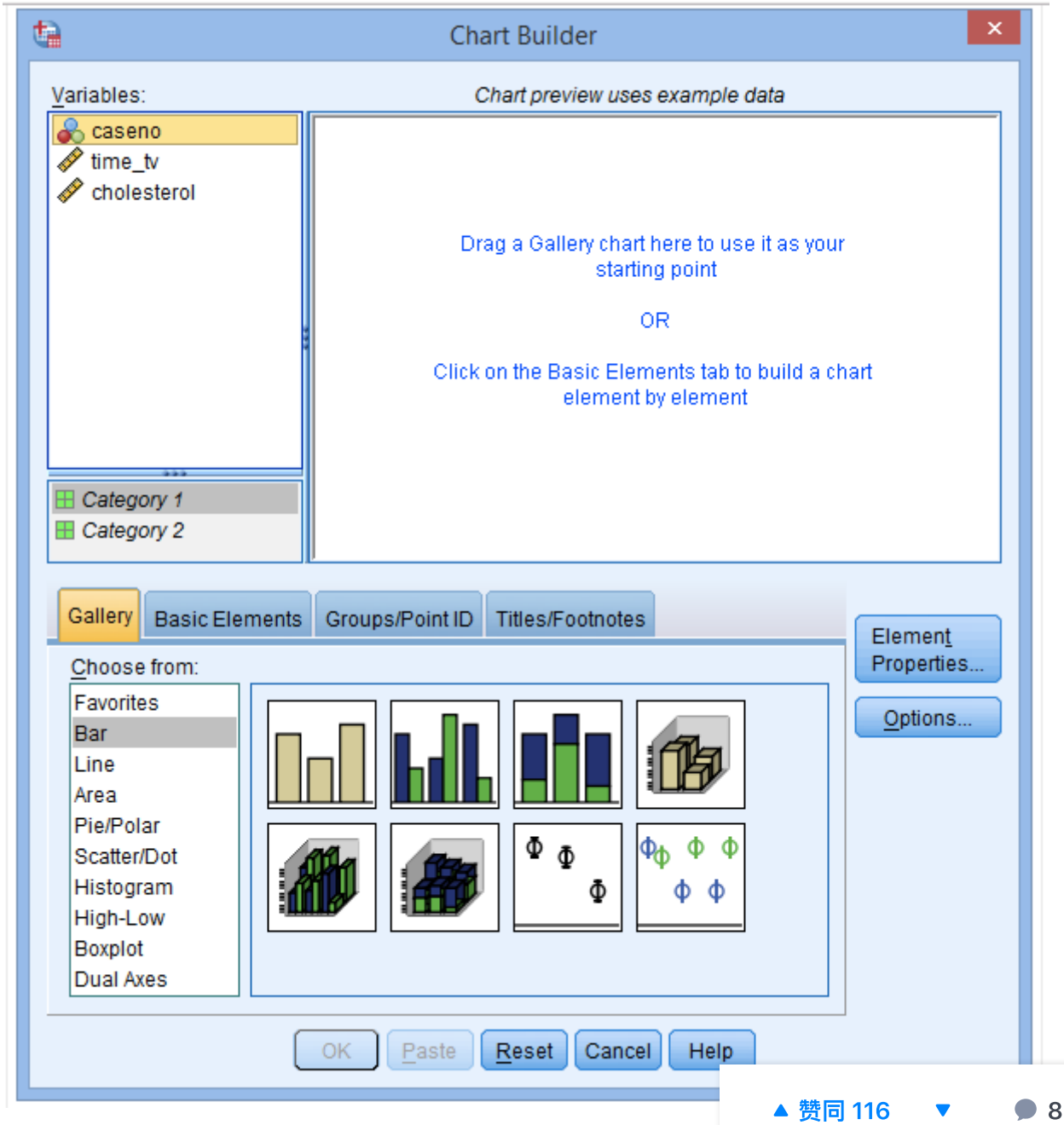
这样的散点图用SPSS怎么画呢?

(1) 在主菜单点击Graphs→Chart Builder

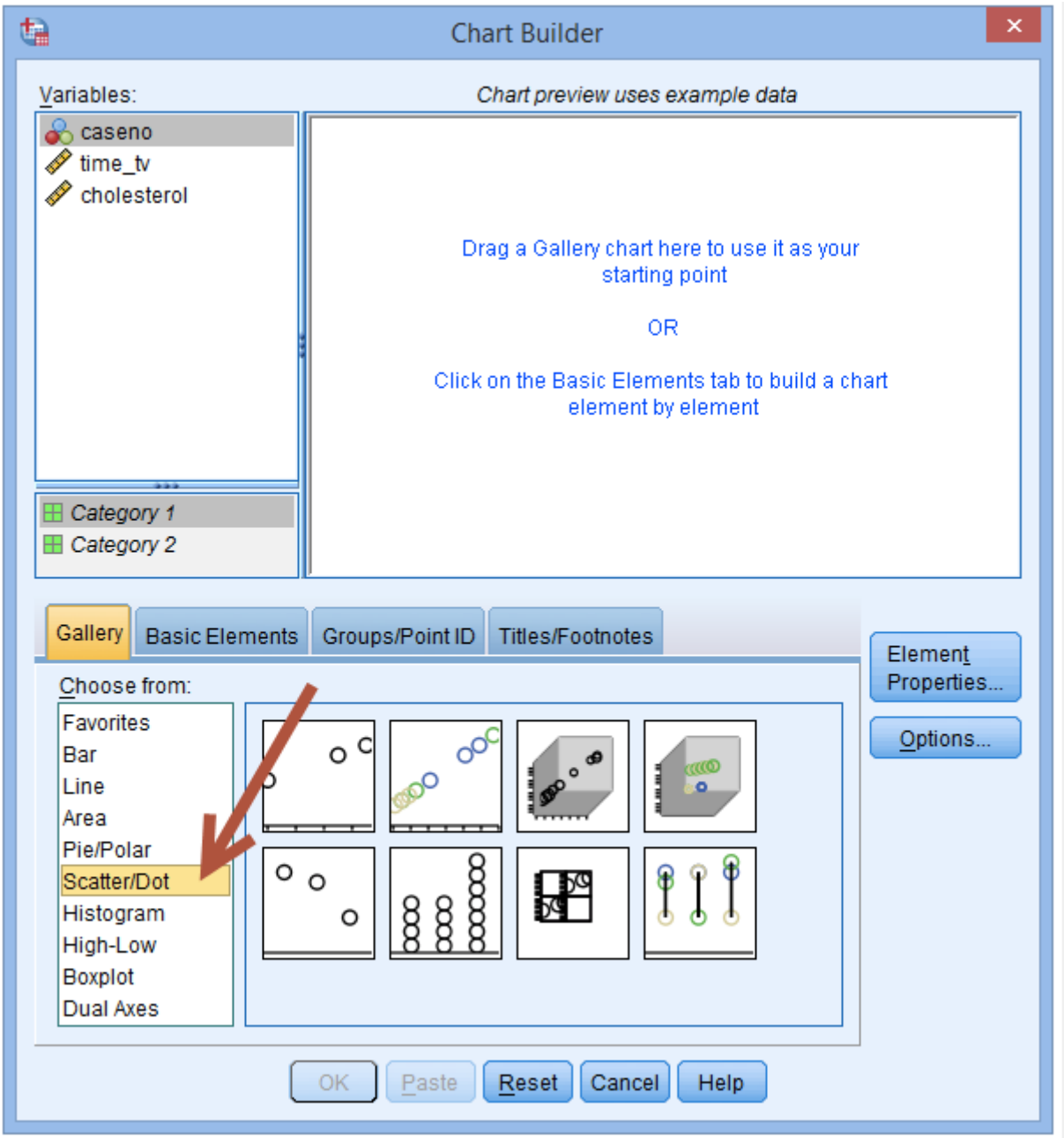


	caseno	time_tv	cholesterol	var	var	var
1	1	168	4.60			
2	2	170	4.80			
3	3	170	5.39			
4	4	164	5.16			
5	5	159	5.09			
6	6	168	5.70			

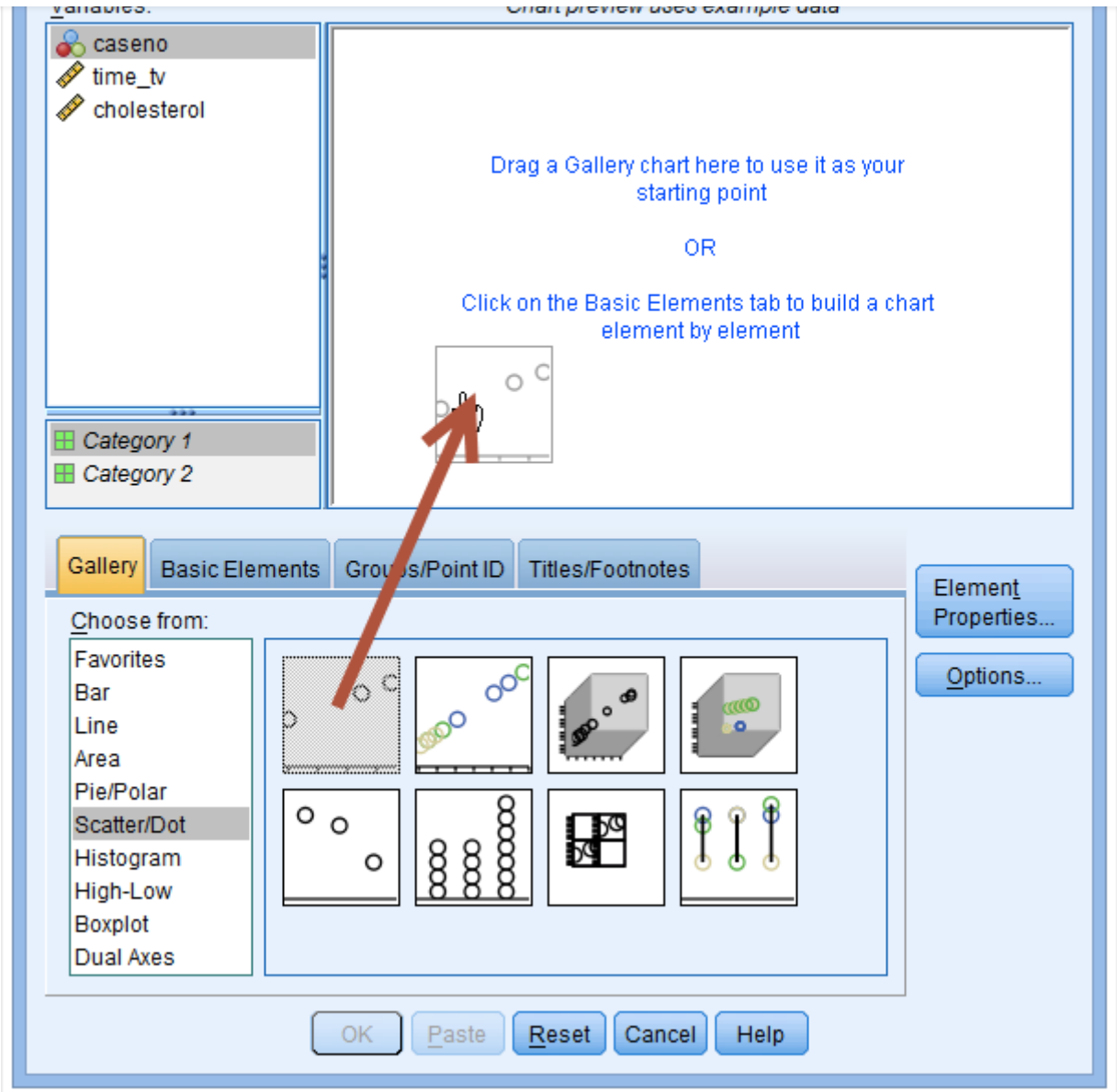
出现下图：



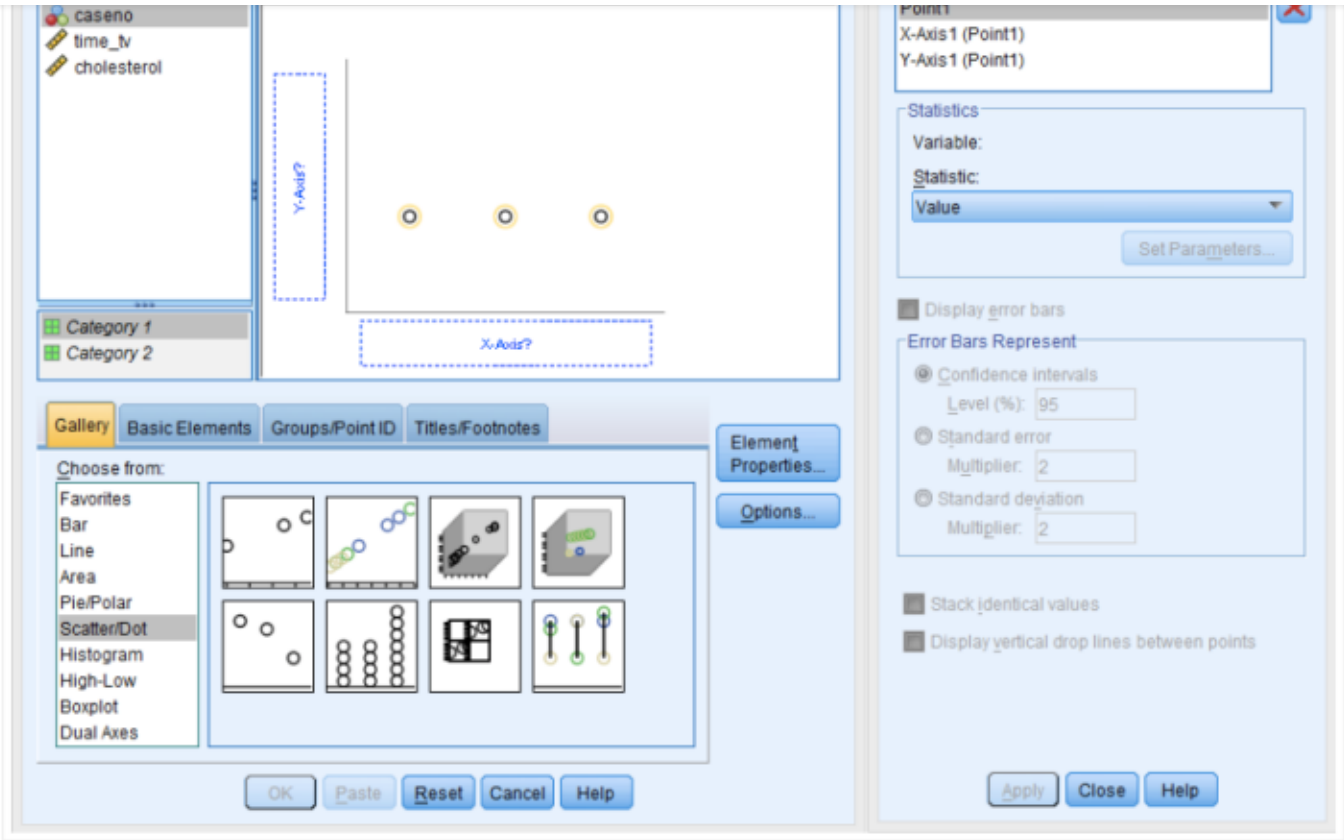
(2) 在Chart Builder对话框下，从Choose from选择Scatter/Dot



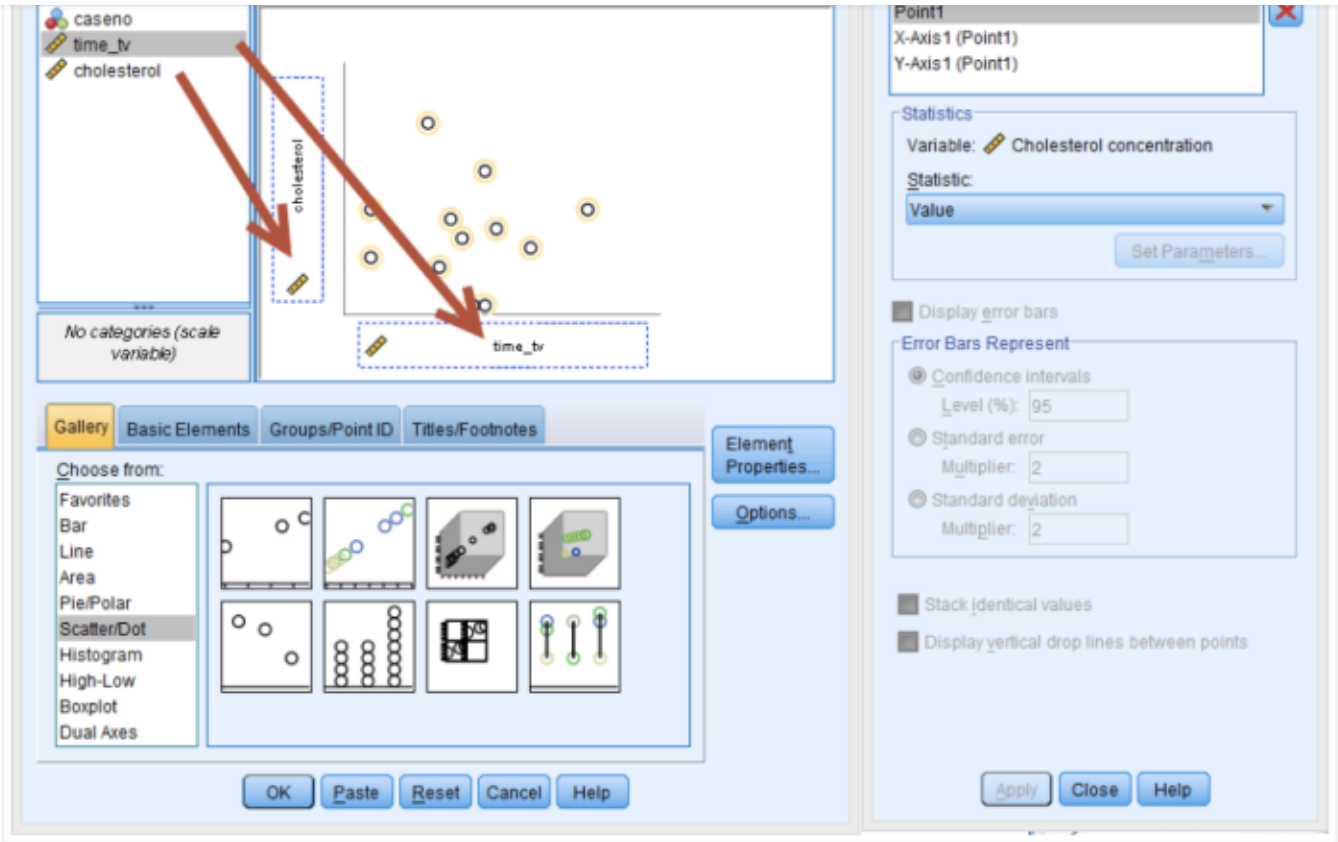
(3) 在中下部的8种图形中，选择左上角的那一种（如果点击这个图标会出现“Simple Scatter”字样），并拖拽到主对话框中



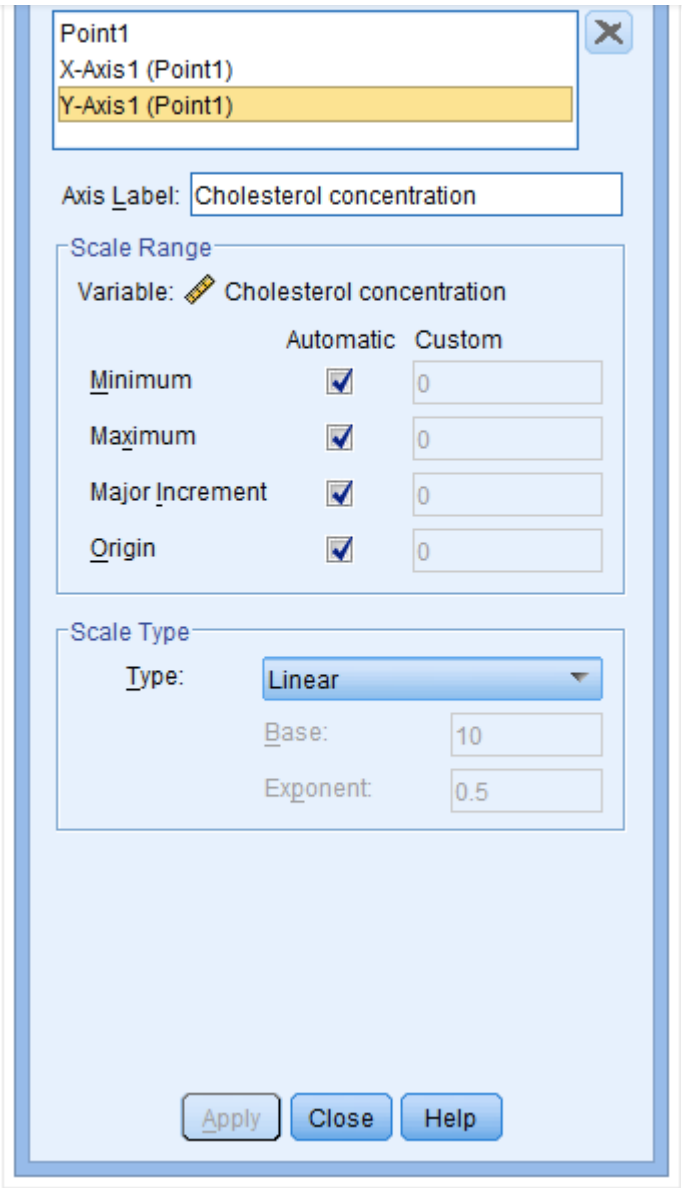
(4) 主对话框中会出现标记“Y-Axis？”和“X-Axis？”的方框



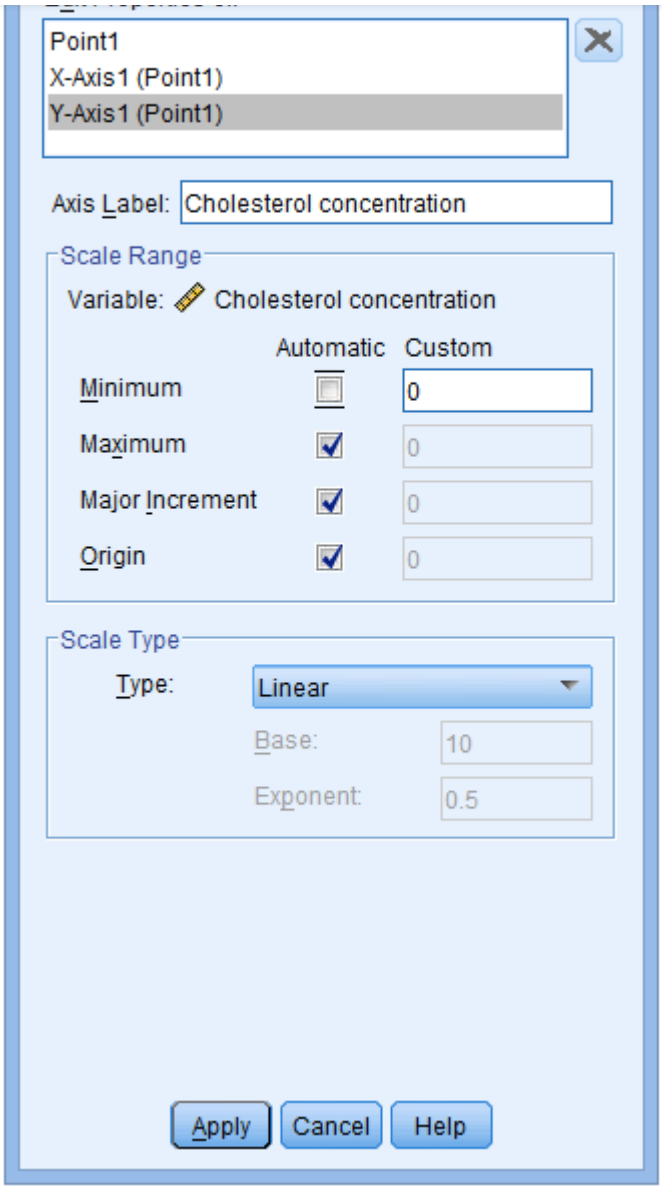
(5) 将看电视时间（time_tv）和胆固醇浓度（cholesterol）变量分别拖拽到“X-Axis？”和“Y-Axis？”方框内



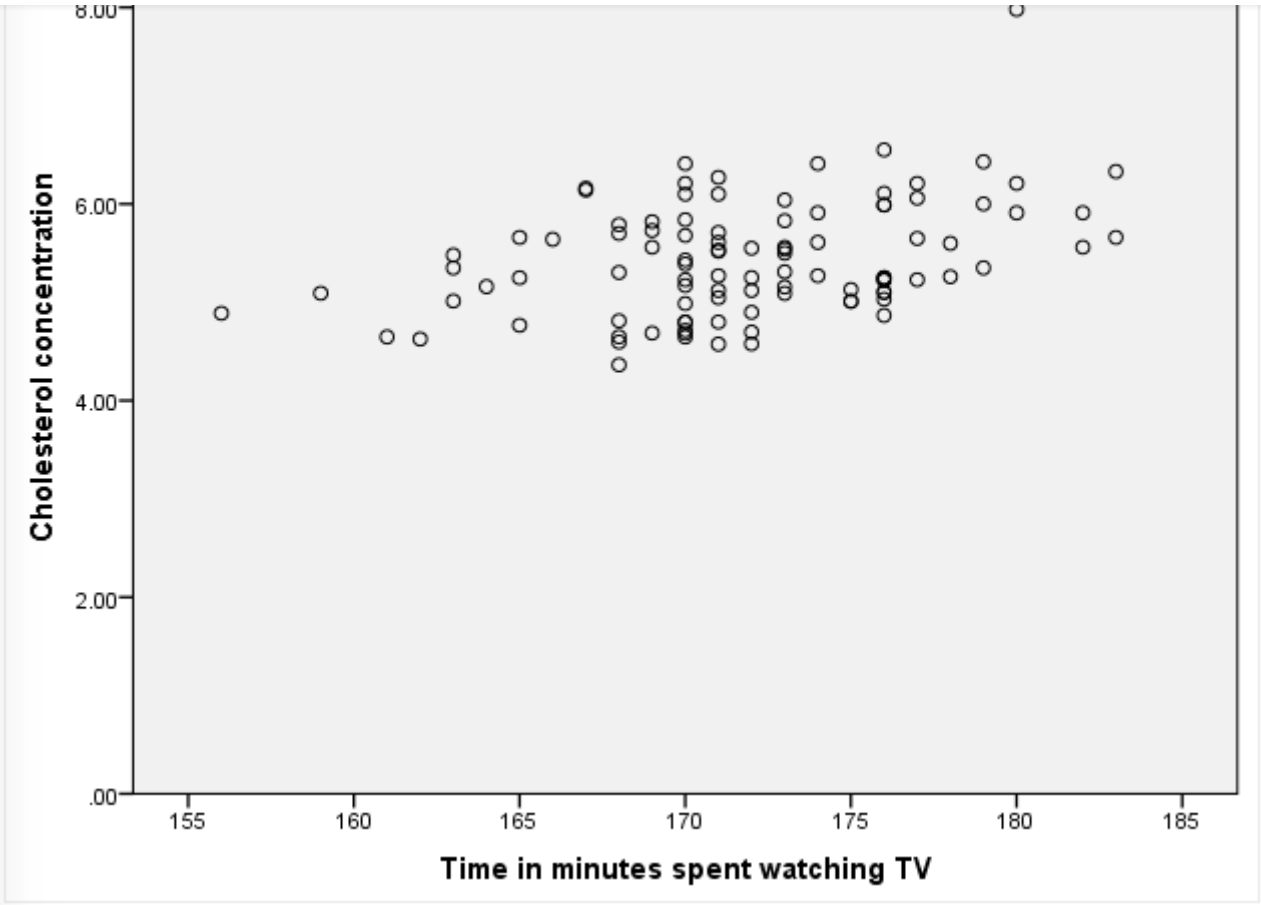
(6) 在Element Properties框内点击Y-Axis1 (Point1)



(7) 在Scale Range框内取消对Minimum的勾选

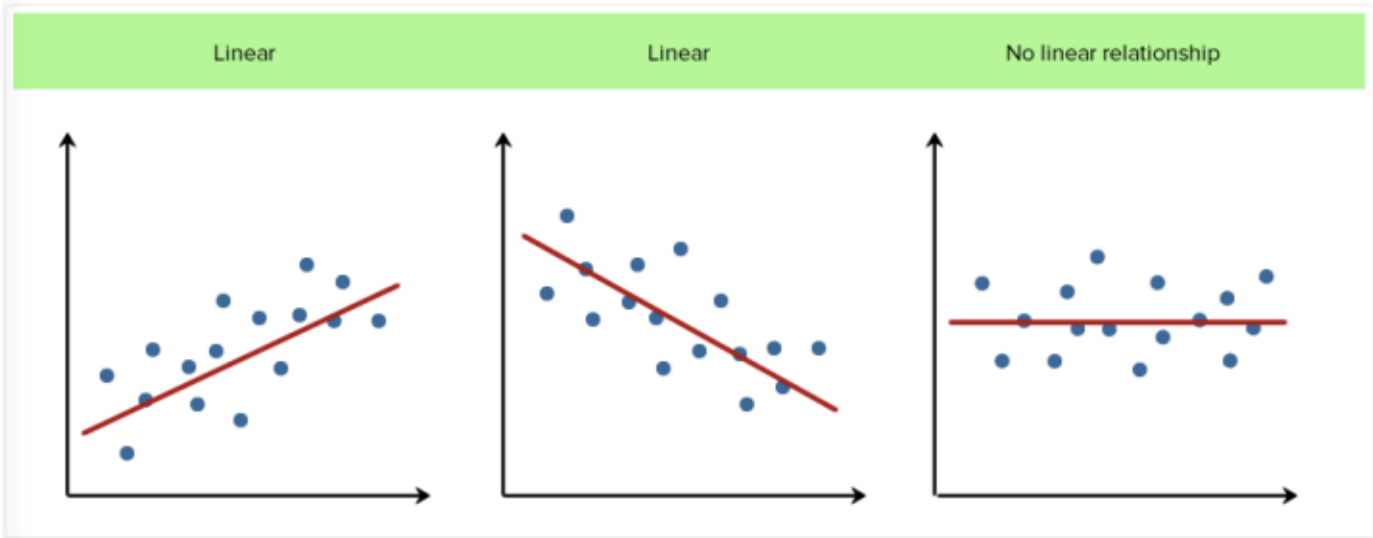


(8) 点击Apply→OK，完成散点图



那么，我们应该如何通过散点图判断是否存在线性关系呢?

我们可以通过简单的视觉判断散点分布是否构成直线，举例如下：



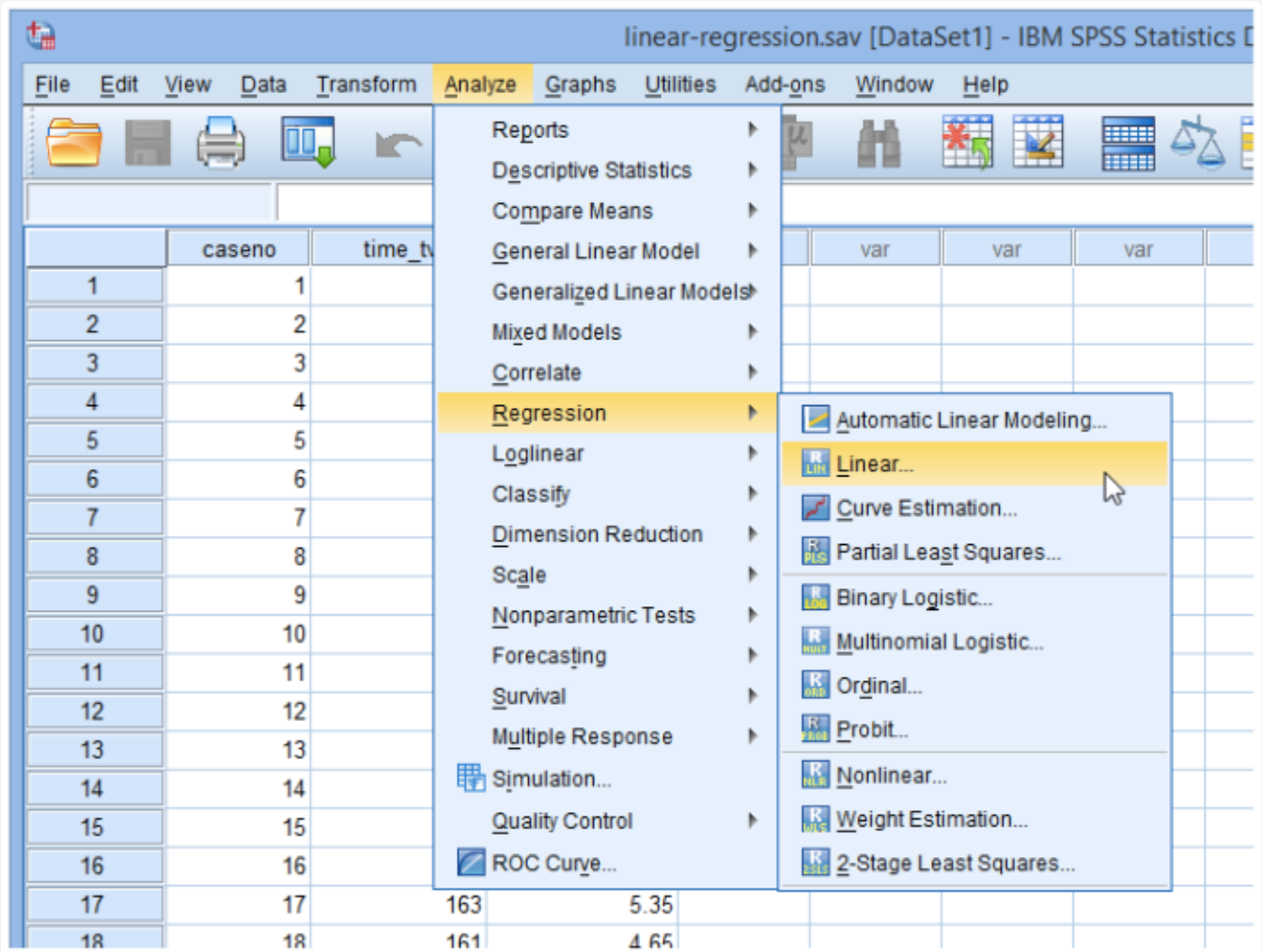
值得注意的是，你可能对右图为什么没有线性关系存在疑问。我们认为简单线性回归中因变量和自变量的线性关系是指因变量会随自变量的变化而发生改变。而虽然右图的散点分布可以构成直线，但是这条直线与X轴平行，证明其因变量不随自变量变化。因此我们认为右图不存在线性关系。

向的，还是负向的，只要因变量和自变量之间存在线性关系，我们就完成了对假设3的检验。

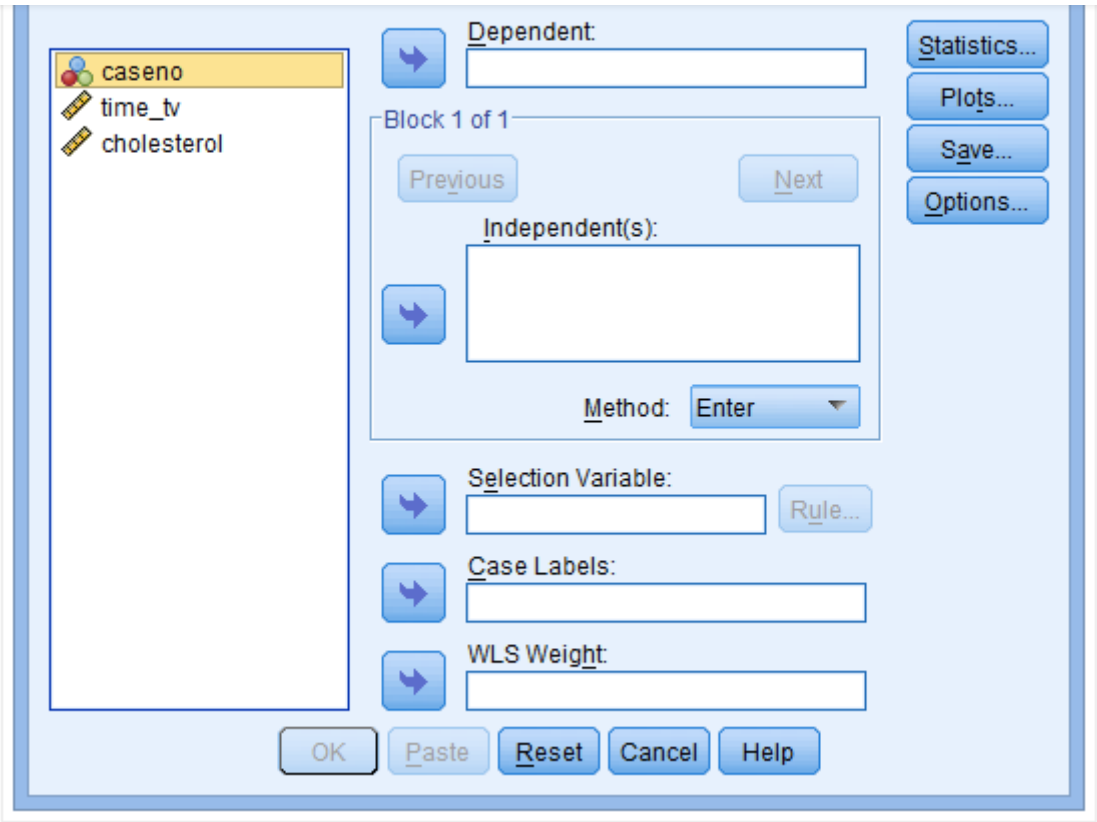
4.3 假设4-7

为了检验假设4-7，我们需要在SPSS中运行简单线性回归，并对结果进行一一分析。

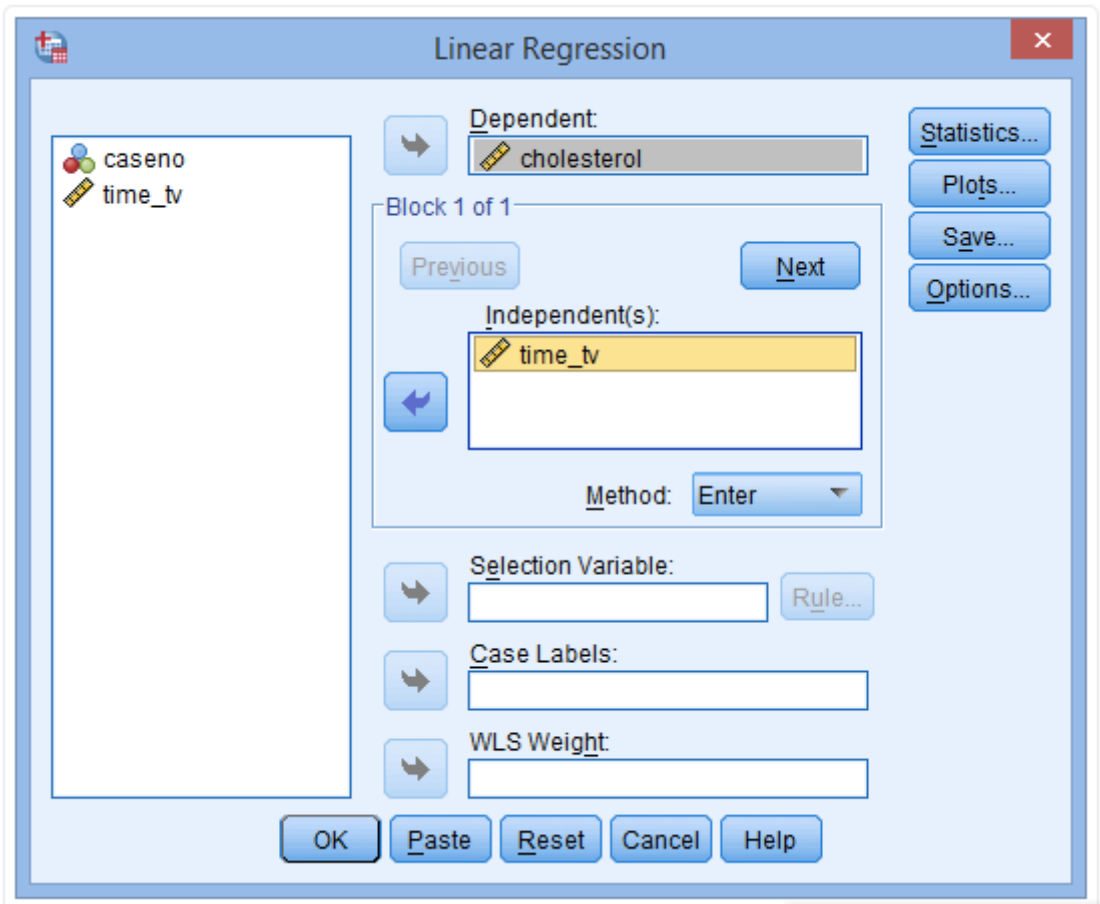
(1) 点击Analyze→ Regression→ Linear



出现下图：



(2) 将看电视时间（time_tv）和胆固醇浓度（cholesterol）分别放入Independent和Dependent栏



(3) 点击Statistics，弹出下图

知乎

(4) 在Regression Coefficient框内点选Confidence intervals，并在Residuals框内点选Durbin-Watson和Casewise diagnosis

(5) 点击Continue，回到主界面

(6) 点击Plots，弹出下图

知乎

(7) 分别在“Y：”和“X：”框内添加“*ZRESID”和“*ZPRED”

(8) 在Standardized Residual Plots中点选Histogram和Normal probability plot

(9) 点击Continue→OK

根据结果，我们将逐一对假设4-7进行检验。

假设4：具有相互独立的观测值

经过上述操作，SPSS输出Durbin-Watson检验结果为：

本研究Durbin-Watson检验值为1.957。一般来说，Durbin-Watson检验值分布在0-4之间，越接近2，观测值相互独立的可能性越大。即，本研究中简单线性回归的观测值具有相互独立性，满足假设4。

但不得不说，Durbin-Watson检验不是万能的。它仅适用于对邻近观测值相关性的检验（1st-order autocorrelation）。举例来说，我们一般按照调查顺序录入数据，将第一位受试者录入到第一行，再将第二位受试者录入到第二行。在这种情况下，Durbin-Watson检验可以检测出第一位受试者和第二位受试者之间的相关性。

但是如果我们乱序录入数据，将第一位受试者和可能与他存在自相关的第二位受试者离得很远，Durbin-Watson检验的结果就不准确了。因此，我们需要慎重对待Durbin-Watson检验的结果。

其实，观测值是否相互独立与研究设计有关。如果研究者确信观测值不会相互影响，我们甚至可以不进行Durbin-Watson检验，直接认定研究满足假设4。

假设5：不存在显著的异常值

在简单线性回归中，异常值是指观测值与预测值相差较大的数据。这些数据不仅影响回归统计，还对残差的变异度和预测值的准确性有负面作用，并阻碍模型的最佳拟合。因此，我们必须充分重视回归的异常值。从看电视时间（time_tv）和胆固醇浓度（cholesterol）的散点图可以看出，本研究存在潜在异常值，如下图标记点：

但是，我们必须注意，由于横纵坐标比例的影响，散点图的直观结果并不可靠。我们需要经过Casewise Diagnostics检验进行客观分析。

经过上述操作，SPSS输出Casewise Diagnostics检验结果为：

结果显示，本研究的第91例数据是潜在异常值，标准残差为4.059。一般来说，Casewise Diagnostics检验标准是上下3倍标准差，并标记超出此范围的数据为潜在异常值。同时，该结果也显示胆固醇浓度的实际值为7.98，而根据潜在异常值预测的胆固醇浓度为5.7977，差值为2.18233。根据这些指标，本研究直接剔除第91例数据，重新进行检验和数据分析。

其实，Casewise Diagnostics检验检测的异常值主要是离群值，如果大家对检测别的异常值感兴趣，可以看我们今后关于杠杆值和影响点的详细介绍。

假设6：等方差性

等方差性是简单线性回归的基本假设，可以通过残差与回归拟合值或标准化残差与标准化预测值之间的散点图进行检验。经过上述操作，SPSS输出结果如下：

如果存在等方差性，不同拟合值对应的残差应大致相同。即图中各点均匀分布，不会出现特殊的分布形状。

如果残差点分布不均匀，形成漏斗或者扇形，那么回归就不具有等方差性，如下图：

(注：increasing funnel, 上升漏斗；decreasing funnel, 下降漏斗；fan shaped, 扇形)

当然，如果研究结果提示不满足等方差性假设，我们也可以通过一些统计手段进行矫正。比如，采用加权最小二乘法进行回归，改用更加稳健的回归或者有稳健标准差结果的回归以及转换数据等。（之后的文章我们会详细介绍~）

假设7：回归残差近似正态分布

(1) 柱状图

经上述操作，SPSS输出结果如下：

从图中可以看出，该回归的标准化残差近似正态分布。但是由于横纵坐标比例的影响，柱状图的结果可能不准确，我们需要绘制正态P-P图进一步验证。

(2) 正态P-P图

正态P-P图各点分布离对角线越近，提示数据越接近于正态分布；如果各点刚好落在对角线上，那么数据就是正态分布。简单线性回归仅要求回归残差接近于正态分布，因此根据上图，我们认为该研究满足假设7。

同时，值得注意的是，相较于柱状图，正态P-P图可以更加明显、准确地判断数据的正态性，具体对比如下。这提示，在判断正态性时，应谨慎对待柱状图的结果，结合正态P-P图进行全面分析。

(注：positive skewness, 正偏；negative skewness, 负偏；positive kurtosis,正峰度；negative kurtosiss, 负峰度；Histogram, 柱状图；normal Q-Q Plot, 正态P-P图)

5、结果解释

简单线性回归可以得到3个主要结果：

(1) 自变量解释因变量变异的比例

(2) 根据新增的自变量预测因变量

(3)自变量改变一个单位，因变量的变化情况

为了更好地解释和报告简单线性回归的结果，我们需要统计以下3个方面：

(1) 线性回归模型的拟合程度

(2) 回归系数

知乎

5.1 判断线性回归模型的拟合程度

判断线性回模型拟合程度的指标有很多，我们主要向大家介绍变异的解释程度、模型的统计学意义以及预测值的准确性（5.3节）3个指标。

5.1.1 变异的解释程度

SPSS简单线性回归输出的结果中有Model Summary表格，如下。其中带有字母“R”的指标（已标黄）与模型对变异的解释程度有关。

第一个标黄的指标R是回归的多重相关系数。当简单线性回归中只有一个自变量时，R值与因变量和自变量的Pearson相关系数相同，代表两者之间的相关程度。如该研究中R=0.359，提示胆固醇浓度与看电视时间中等相关。但实际上，简单线性回归并不关注R值。

第二个标黄的指标R2（R Square）代表回归模型中自变量对因变量变异的解释程度，是分析回归结果的开始。本研究中，R2=0.129，提示自变量（看电视时间）可以解释12.9%的因变量（胆固醇浓度）变异。但是，R2是基于样本数据计算出来的，会夸大自变量对因变量变异的解释程度。

第三个标黄的指标adjusted R2（Adjusted R Square）。与R2不同的是，它剔除了自变量个数的影响，准确性更好。本研究中，adjusted R2=0.120，小于R2=0.129，校正了R2对总体自变量对因变量变异解释程度的夸大作用。同时，adjusted R2也是影响程度的评价指标。本研究中，adjusted R2=0.120，提示中等影响。

5.1.2 模型的统计学意义

SPSS的输出结果中有ANOVA表格，如下：

该表中各指标的含义如下：

结果显示，本研究回归模型具有统计学意义，F(1, 97)=14.39，P<0.001，提示因变量和自变量之间存在线性相关。如果P>0.05,则说明该回归没有统计学意义，因变量和自变量之间不存在线性相关。

5.2 回归系数的解释

本研究的回归方程可以表示为：

cholesterol= b0+(b1×time_tv)

其中，b0是截距，b1是斜率。如果可以得到这两个指标，我们就可以根据自变量（看电视时间，time_tv）预测因变量（胆固醇浓度，cholesterol）了。SPSS对回归截距和斜率的输出结果如下：

在SPSS中，截距被称为“Constant”，即-0.944，如下：

知乎

实际上，我们并不是关注回归的截距指标。它是指当自变量为0时，因变量的值。在本研究中，回归截距提示当看电视时间为0，即从来不看电视时，受调查者胆固醇浓度的平均值为-0.944mmol/L。这种分析方法是不对的，不仅因为它不符合客观实际，还因为它存在对数据过度挖掘的风险。同时，我们也可以通过P值判断截距的统计学意义，如下：

通过P值（P=0.575），我们也可以看出该研究的截距没有统计学意义，即截距值（-0.944）与0的差异没有统计学意义。必须强调的是，无论截距的统计检验结果如何，我们在进行简单线性回归时都不是十分关注这项指标。我们主要的关注指标是斜率，如下标黄的部分：

斜率代表的是自变量每改变一个单位因变量的变化值。在本研究中，看电视时间的斜率是0.037，表示每当看电视时间增加1分钟，胆固醇浓度增加 0.037mmol/L。

举例来说，如果某受调查者看电视时间从170分钟/天增加到180分钟/天（增加10分钟/天），她/他的胆固醇浓度将增加0.037×10=0.370mmol/L。同样地，我们也可以计算出每当看电视时间增加5、15、20分钟/天时，对应胆固醇浓度的增加值。但是，我们并不能无限制地改变看电视时间。

为了避免对数据的过度挖掘，我们一般要求在自变量观测到的最大值和最小值之间进行计算。

根据SPSS结果，我们也可以得到斜率的可能范围，如下标黄的部分：

从表中可以看出，斜率的95%置信区间在0.018-0.056mmol/L（Lower Bound，Upper Bound）。同时，在Sig栏可以得到斜率的统计学检验结果，如下：

斜率的P值为0.000（在报告中应记为P<0.001），提示斜率值与0的差异有统计学意义，也说明胆固醇浓度与看电视时间存在线性关系。

如果斜率的P值大于0.05，证明斜率没有统计学意义，即斜率值与0的差异没有统计学意义，说明因变量和自变量之间不存在线性关系。在这种情况下，我们不能通过自变量预测因变量。

将系数代入回归方程，得：

cholesterol= -0.944+(0.037×time_tv)

根据这个方程，我们可以计算合理范围内任意看电视时间对应的胆固醇浓度。但针对这个例子，仅依靠看电视时间计算胆固醇浓度存在专业上的质疑。因此，我们仅认为看电视时间是久坐生活习惯的一项指标，通过该模型可以对胆固醇浓度做出一些解释。

5.3 预测因变量

简单线性回归的一个主要作用就是根据自变量预测因变量。正如5.2提到的，我们仅根据看电视时间预测胆固醇浓度存在专业质疑，但是为了系统地向大家介绍简单线性回归的功能，我们仍用这个例子进行讲解。

这一节，我们从根据回归方程预测因变量开始，逐步向大家介绍计算预测值和95%置信区间的SPSS操作方法及对预测结果的解释。

5.3.1 根据回归方程计算预测值

根据SPSS结果，我们得到本研究的线性回归方程如下：

我们仅需要将看电视时间代入方程就可以得到胆固醇浓度的预测值。举例来说，如果某位受试者每天看电视的时间为180分钟（3小时），带入方程如下：

预测的胆固醇浓度= -0.944+(0.037×180)=5.72 mmol/L

即，当看电视时间为180分钟/天时，预测胆固醇浓度为5.72 mmol/L。

这个预测值有两种含义。第一，如果我们调查了目标人群中所有电视时间为180分钟/天的人，他们胆固醇浓度的平均值应为5.72 mmol/L。第二，如果某位受调查者看电视的时间为180分钟/天，那么5.72 mmol/L是其胆固醇浓度的最佳估计值。

第二种含义比较难理解，在此我们向大家具体说明一下。大家都知道，即使两个人看电视的时间相同，他们实际的胆固醇浓度也可能不同。我们用平均值描述他们的情况比用某一个人的实际值好。因此，我们认为用看电视时间为180分钟/天的受调查者胆固醇浓度的平均值代表这个群体更好，即5.72 mmol/L是其胆固醇浓度的最佳估计值。

5.3.2 预测值和95%置信区间的SPSS操作方法

相较于5.3.1的计算方法，SPSS操作可以同时进行多个数据的计算，并估计预测值的95%置信区间。我们以看电视时间为160、170和180分钟/天为例，向大家介绍预测值和95%置信区间的SPSS操作方法。

(1) 点击Analyze→ General Linear Model→ Univariate，出现下图：

(2) 将因变量cholesterol放入Dependent Variable框内，自变量time_tv放入Covariate(s)框内

(3) 点击Paste，出现IBM SPSS Statistics Syntax Editor窗口如下：

(4) 在/DESIGN=time_tv.上方插入/LMATRIX=ALL 1 160，如下：

语法解释：在只有一个自变量的简单线性回归中，LMATRIX命令允许加入自变量的数值。/LMATRIX=ALL 1 160语句中各部分的含义如下：

ALL指同时运用斜率和自变量进行预测；

1 指纳入截距；

160 指用来预测因变量的自变量值。

如果我们想同时进行多组预测，只需要在该语句后面加";ALL 1 VALUE"。其中，VALUE是指用于预测因变量的自变量值。例如，我们要预测看电视时间为160、170和180分钟/天时的胆固醇浓度，如下：

(5) 点击Run→ All，输出结果

5.3.3 预测结果的解释

预测结果在Contrast Results（K Matricx）中展示，如下：

知乎

我们是以看电视时间为160、170和180分钟/天为例进行预测的，语法是

LMATRIX=ALL 1 160; ALL 1 170; ALL 1 180

结果也是按照语法顺序进行排列的，即L1（红框）是每天看电视时间为160分钟的预测值，L2 是（蓝框）是每天看电视时间为170分钟的预测值，L3 是（绿框）是每天看电视时间为180分钟的预测值。

我们以每天看电视时间为160分钟为例解释预测结果，如下图红框部分：

从Contrast Estimate可以看出，每天看电视160分钟的胆固醇浓度预测值为4.98 mmol/L。我们根据回归方程可以得到相同的结果 -0.944+0.037×160 =4.98 mmol/L。但是，SPSS操作还提供了其他结果。如，预测值的标准误（Std. Error）是0.13 mmol/L，提示预测值的变异程度。再如，预测值的95%置信区间（Confidence Internal for Difference）为4.73-5.23 mmol/L

也可以使用回归方程进行计算，但是我们得到的区间估计不是置信区间，而是预测区间。由于个体观测值的不稳定性，预测区间往往比置信区间大。同时，个体预测的预测区间不能通过SPSS自动计算得到。在本章节，我们只需要记得个体预测的预测区间与样本预测的置信区间不同即可。

6、撰写结论

6.1 简洁汇报

简单线性回归结果提示，看电视时间与胆固醇浓度之间存在线性关系F(1,97) = 14.395（P <0.001）；看电视时间可以解释胆固醇浓度变异的12.9%。回归方程如下：

胆固醇浓度= -0.944+(0.037×看电视时间)

6.2 统计结果报告

采用简单线性回归模型分析看电视时间对胆固醇浓度的影响。通过绘制散点图，直观判断两者之间存在线性关系，并通过绘制标准化残差散点图和带正态曲线的柱状图或P-P图，验证数据具有等方差性和残差正态性。同时为了保证数据的代表性，我们剔除了一项异常值（胆固醇浓度为7.98 mmol/L）。回归方程如下：

胆固醇浓度= -0.944+(0.037×看电视时间)

看电视时间对胆固醇浓度的影响有统计学意义，F(1,97)=14.395（P <0.001）；看电视时间可以解释胆固醇浓度变异的12.9%，影响程度中等（调整R2= 12.0%）。每增加1分钟/天看电视时间，胆固醇浓度增加0.037 (95% CI：0.018-0.056)mmol/L。此外，看电视时间为160分钟/天、170分钟/天和180分钟/天的胆固醇浓度预测值分别为4.98 (95% CI：4.73-5.23)mmol/L、5.35 (95% CI：5.24-5.45)mmol/L和5.72 (95% CI：5.53-5.90)mmol/L。

6.3 散点图

根据4.2的讲解，我们已经可以绘制出基本的散点图，如下：

知乎

但是在汇报结果时，我们仍需要增加最佳拟合线、置信区间和预测区间等指标。具体操作方法如下：

(1) 双击散点图，激活Chart Editor

(2) 点击Element→ Fit Line at Total

出现下图：

同时，Properties对话框也会自动弹出

知乎

提示：如果只想做出最佳拟合线，到这一步就可以关闭Properties和Chart Editor窗口，Output Viewer窗口会自动出现下图，完成操作。

如果需要绘制置信区间和预测区间，请继续第(3)步的操作。

(3) 在Properties对话框中，点击Confidence Intervals中的Mean

(4) 点击Apply，出现下图

(5) 在Properties对话框中，点击Confidence Intervals中的Individual

(6) 点击Apply，出现下图

(7) 关闭Properties和Chart Editor窗口，Output Viewer窗口会弹出带有置信区间和预测区间的散点图

知乎

(8) 但是，一般学术报告都要求去除背景和边框颜色，这应该怎么做呢？双击散点图，激活properties窗口，在Fill & Border窗口内修改背景颜色

(10) 点击Apply，背景颜色从灰色变为无色

(12) 点击Apply，边框颜色从黑色变为无色，图中上方和后侧的边框线消失

(13) 关闭Properties窗口

(14) 那如果想改变坐标轴数字的保留位数，应如何做呢？双击纵坐标轴上的任意数字（如6.00），激活纵坐标轴的Properties窗口

(15) 点击Properties窗口内的Number Format

(16) 将Decimal Places框内的“2”改成“1”

(17) 点击Apply，纵坐标数据由保留两位小数变为保留一位小数。但实际上，在本研究中胆固醇浓度保留两位小数比较合理，所以我们仍保留两位小数

(18) 点击Close，关闭Properties窗口

(19) 再进一步调整线型后，我们就可以得到学术出版要求的散点图，如下

知乎

7、延伸阅读

简单线性回归异常值的处理

数据异常值主要有以下三类：

(1) 数据录入错误

当出现异常值时，首先应考虑是否存在录入错误。这是最简单的异常值类型，我们只需要查到原数据，重新录入即可。

(2) 数据测量错误

如果不存在录入错误，我们就需要检查异常值是不是由测量错误导致的。比如，用量程为0-100°C的测试仪器测量温度，结果发现有些数据超过100°C，那么我们就推测这些数据是由于测量错误导致的。

在大多数情况下，测量错误都无法弥补，我们一般建议直接剔除这些数据。但如果我们知道这些异常值的方向，如上述的例子中，存在大于100°C的数据，我们可以录入为上限值100°C。虽然这样会造成偏倚，但对数据的影响仍小于直接剔除异常值。

如果异常值既不是录入错误，也不是测量错误，是数据中自然存在的，那么我们就不能仅仅因为这些异常值影响了线性回归的基本假设就直接剔除。针对这类异常值，既往研究没有统一的处理意见，建议研究者按照自己的喜好处理。

处理异常值后需要重新进行检验和分析。同时，值得注意的是，如果数据中存在多于一个异常值，我们可以先处理其中比较严重的，并重新检验，可能其他潜在异常值就不再是异常值了。

发现异常值后，我们如何做呢?

7.1 保留异常值

如果不希望或者不能剔除异常值，我们可以采取以下措施：

(1) 对因变量进行数据转换

数据转换可以改变数据的分布比例，从而影响异常值的检验结果。但由于数据转换，回归系数会比较难解释，增加了数据分析的难度。同时，我们也必须确定转换后的数据满足等方差性和残差正态性，重新检验回归假设。

(2) 分别运行纳入和不纳入异常值的回归模型，若结果没有差异，保留异常值

剔除或者处理异常值的目的是为了减小异常值对回归结果的影响。如果能证明数据中的异常值对回归结果（如回归系数和置信区间）没有明显影响，我们就可以保留异常值。即分别运行纳入和不纳入异常值的回归对比结果，分析异常值对回归结果的影响程度，从而判断异常值的去留。

(3) 选择更稳健的回归模型

我们也可以通过调整标准误，运行更稳健的回归模型，但是SPSS现在还没有这项操作。

7.2 剔除异常值

我们可以直接剔除异常值，但这往往是我们迫不得已的做法。因为我们进行数据分析是为了根据样本结果推论总体，但直接剔除异常值就相当于不再考虑这部分人的信息，忽略了他们在总体人群中的作用。

如果一定要剔除异常值，我们就应该在报告中描述被剔除者的信息（数据以及对研究结果的影响）。这样读者就可以清楚地了解到我们剔除异常值的原因以及这

举例来说，本研究中异常值的胆固醇浓度为7.98 mmol/L，远高于普通人群的胆固醇浓度，提示存在心脏病风险。尽管我们希望了解人群胆固醇浓度的基本情况，但是我们并不想纳入存在临床指征或心脏病高危风险的患者。胆固醇浓度这么高的人不是我们的目标人群，所以本研究直接剔除该异常值。

（更多内容可关注“医咖会”微信公众号：传播医学知识和研究进展，探讨临床研究方法学。）

编辑于 2017-05-26

统计　医学　医学统计

推荐阅读

多重线性回归的结果解读和报告（SPSS实例教程）

我们推送了“多重线性回归的SPSS详细操作步骤”，介绍了在应用多重线性回归模型之前所需要满足的8个适用条件，简单概括如下：(1) 自变量与因变量存在线性关系；(2) 残差间相互独立；(3) 残...

医小咖

SPSS实例教程 对照Logistic回

1、问题与数据某医室的患者数据，采照方法探究吸烟和研究为每一位肺癌（±2岁）、性别和配2名对照，对病例

医小咖

8 条评论

切换为时间排序

写下你的评论...

初九未成　1 年前

写的真好，我花了一下午时间看

赞

解决了很多具体操作的问题，很实用

1

gloria 6 个月前

超级实用，牛！

1

Tracy 4 个月前

正态性检验 可以用残差的Shapiro-Wilk检验吗?

赞

丁榕 4 个月前

太良心了！！海外留学狗点开微信公众号发现是贵P小伙伴的时候简直要哭出来惹

赞

jessica gao 3 个月前

太牛了~我都看烦了 是怎么写出来的

1

jessica gao 3 个月前

有个问题哎 为什么在添加总拟合线的时候 按钮是灰色的?

赞

Para 2 天前

你好题主~想问下我的只能单独选择置信区间或者预测区间是怎么回事呢？最后只能输出一种呀。。

赞