解密天猫双11成交数据的“神奇”拟合：统计的障眼法

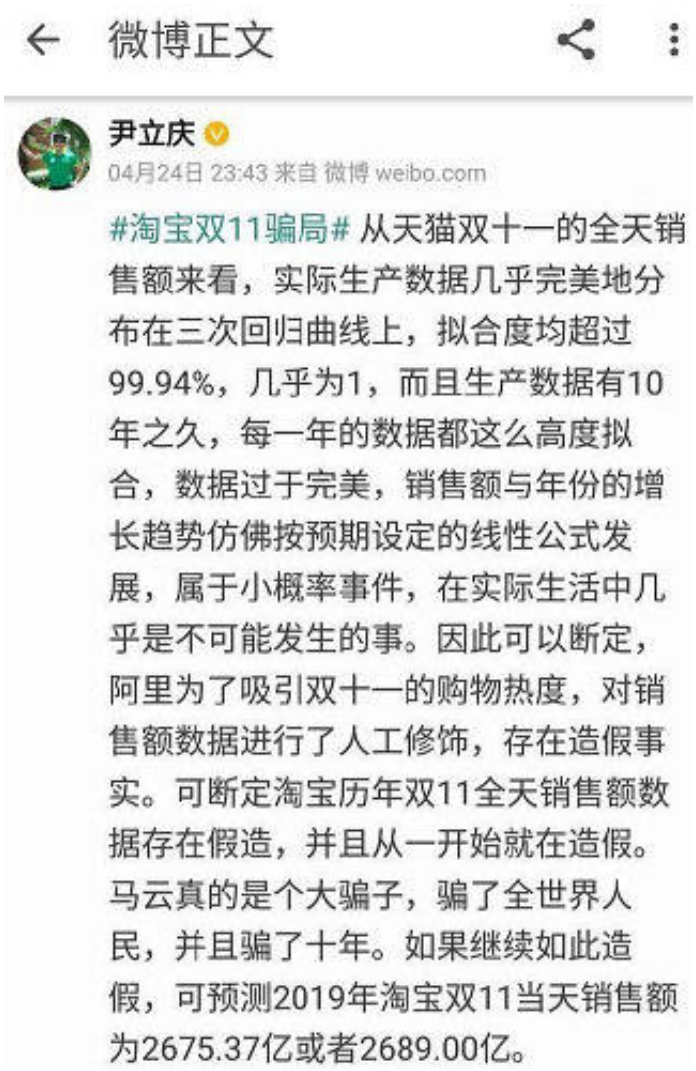
风云学会陈经 11.13 12:08 阅读 75524

不要相信拟合的神奇，也不要相信“拟合度”接近1的神奇效果。这在数学统计里，实在很平常。

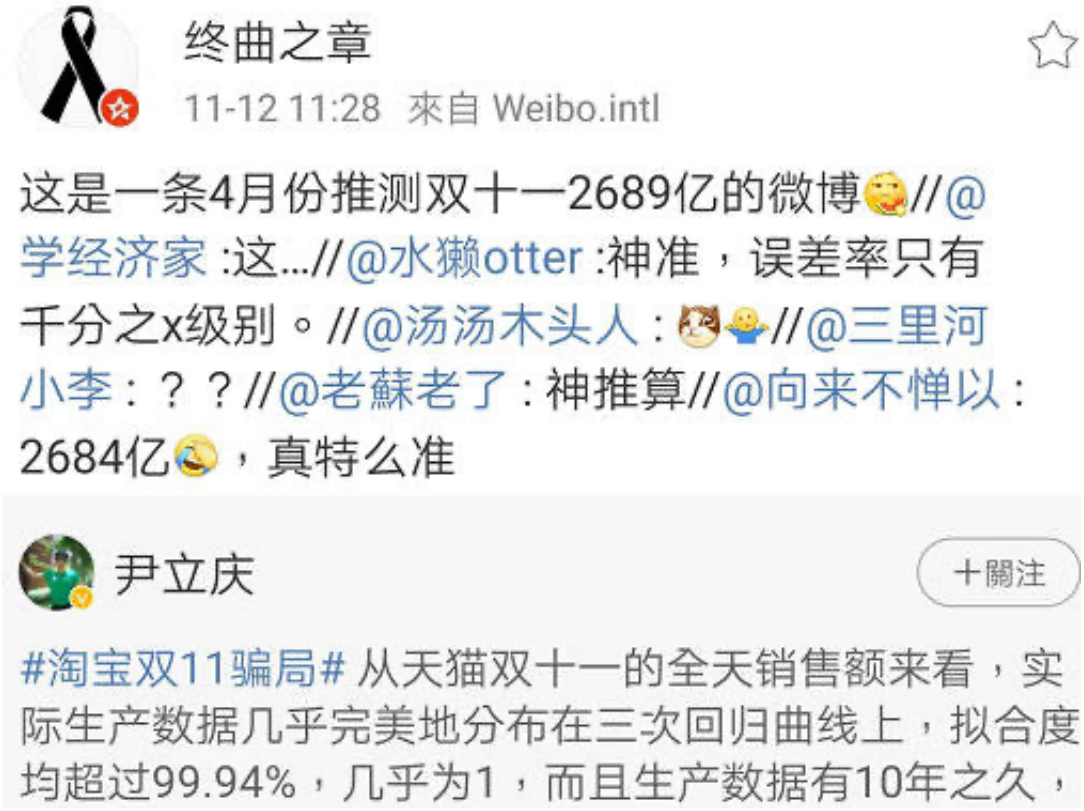
陈经
2019年11月13日
（一会发风云之声）

2019年双11过去以后，阿里报告了天猫双十一的全天销售额：2684亿。这个数字引发了一场关于统计学的有趣争执。

一位叫“尹立庆”的微博网友，在2019年4月24日发了一个贴，通过拟合2009–2018年的双11天猫数据，由于拟合度高达99.94%，他认为淘宝是在按公式假造成交数据。并且他还“预测”，天猫2019年双11成交额为2675.37亿（二次拟合）或者2689亿（三次拟合）。最终出来的数据是2684亿，与他预测的2689亿非常接近。



这个“精准”的提前预测引发了不少人的关注，很多人确实相信淘宝是在凑成交额，不然怎么可能这么准？



很多人翻出了这个“神预测”，暗示“阿里数字造假被抓现行”。这引发了一场风波，许多人在传，阿里双11销售数字造假。尹立庆在微博上的原贴已经被删除，但是网上截图还是很多。

天猫的公关负责人也发了声明，驳斥了造假的说法。

彭美天猫公关总监

早上醒来，有朋友发给我一条据说【强大的预测】。

这种预测看起来真的很"唬人"阿。按照这位网友的逻辑，符合统计趋势的就是假的。那么，世界经济总量也是能被预测的，经济发展也是假的吗？

自己YY下满足自嗨就算了，由此得出天猫双11数据造假，就是造谣了哦，要负法律责任的！

12日晚间，天猫再度发文回应称，“今早到现在，这则精心图文化设计的‘预测’开始被刻意传播”、“已就这则谣言启动司法流程”。估计尹立庆删微博是觉得不太对了。但是个人感觉天猫的回应没有从数学上解释这些疑问。

到底阿里有没有对双11天猫成交数据造假？如果没有造假，为什么统计拟合如此精准，尹立庆提前半年的预测又如此准确？我们来介绍一下相关的知识。

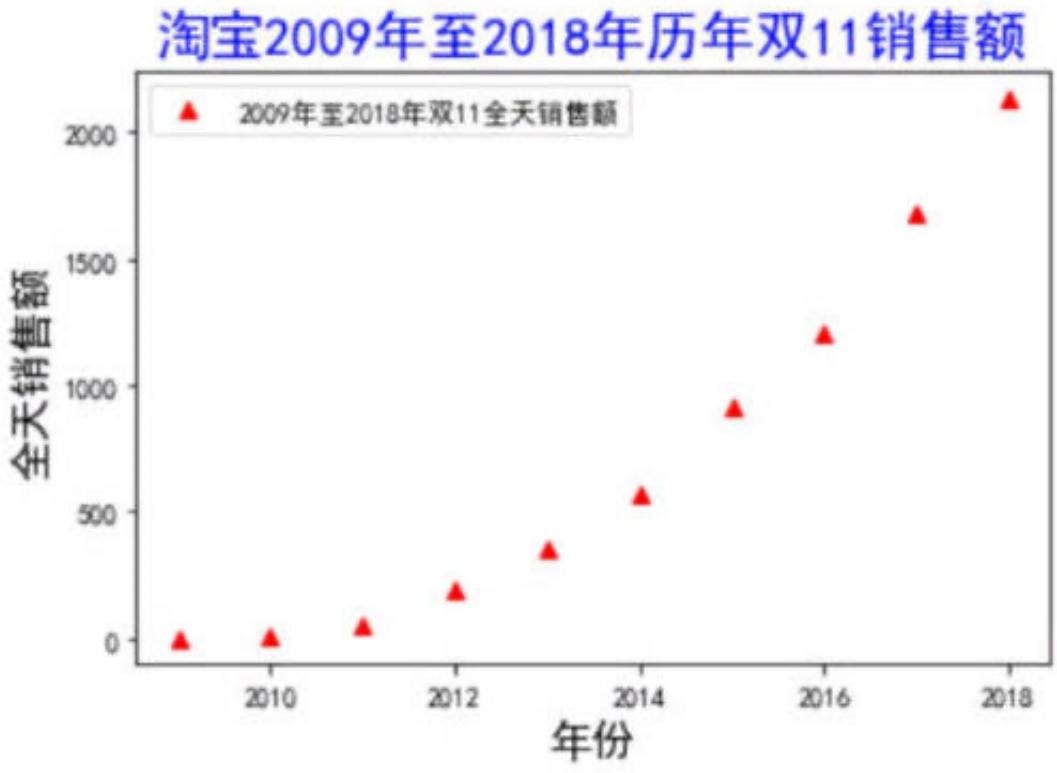
首先说一下，这个预测涉及的“二次拟合”或者“三次拟合”不需要手算，其实是Excel等数据表格软件的功能。所以，不需要进行高深的数学推理和计算，会用Excel简单地制表就行了。

首先是数据源，这个是有公开数据的，没有争议：

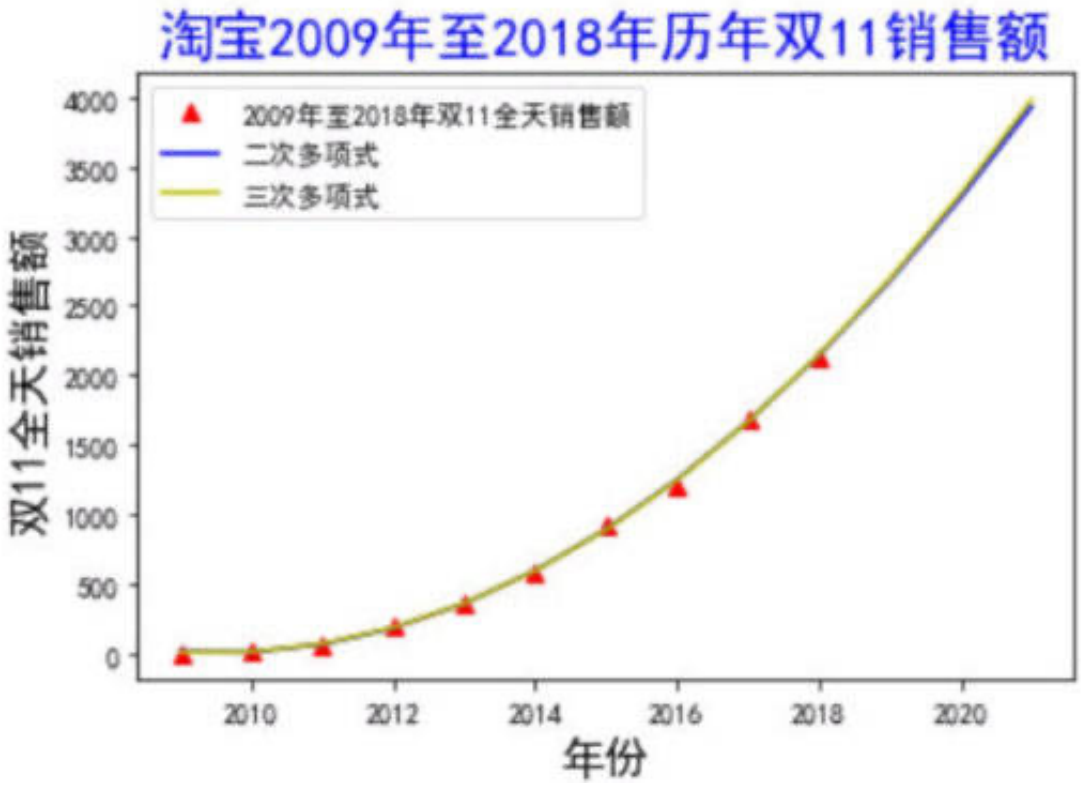
淘宝双11销售额.txt

	年份	淘宝双11销售额
2	2009	0.50
3	2010	9.36
4	2011	52.00
5	2012	191.00
6	2013	350.00
7	2014	571.00
8	2015	912.00
9	2016	1207.00
10	2017	1682.69
11	2018	2135.00
12		

然后，对这两列数据制一个散点图表：



然后是二次和三次多项式拟合曲线。



尹立庆的关键预测是下面这个：

二次多项式拟合 r-squared：0.999377665275

采用二次多项式预测未来三年造假结果：

预测 2019 年淘宝双 11 当天销售额：2675.37 亿

预测 2020 年淘宝双 11 当天销售额：3273.00 亿

预测 2021 年淘宝双 11 当天销售额：3930.76 亿

三次多项式拟合 r-squared：0.999392944966

采用三次多项式预测未来三年造假结果：

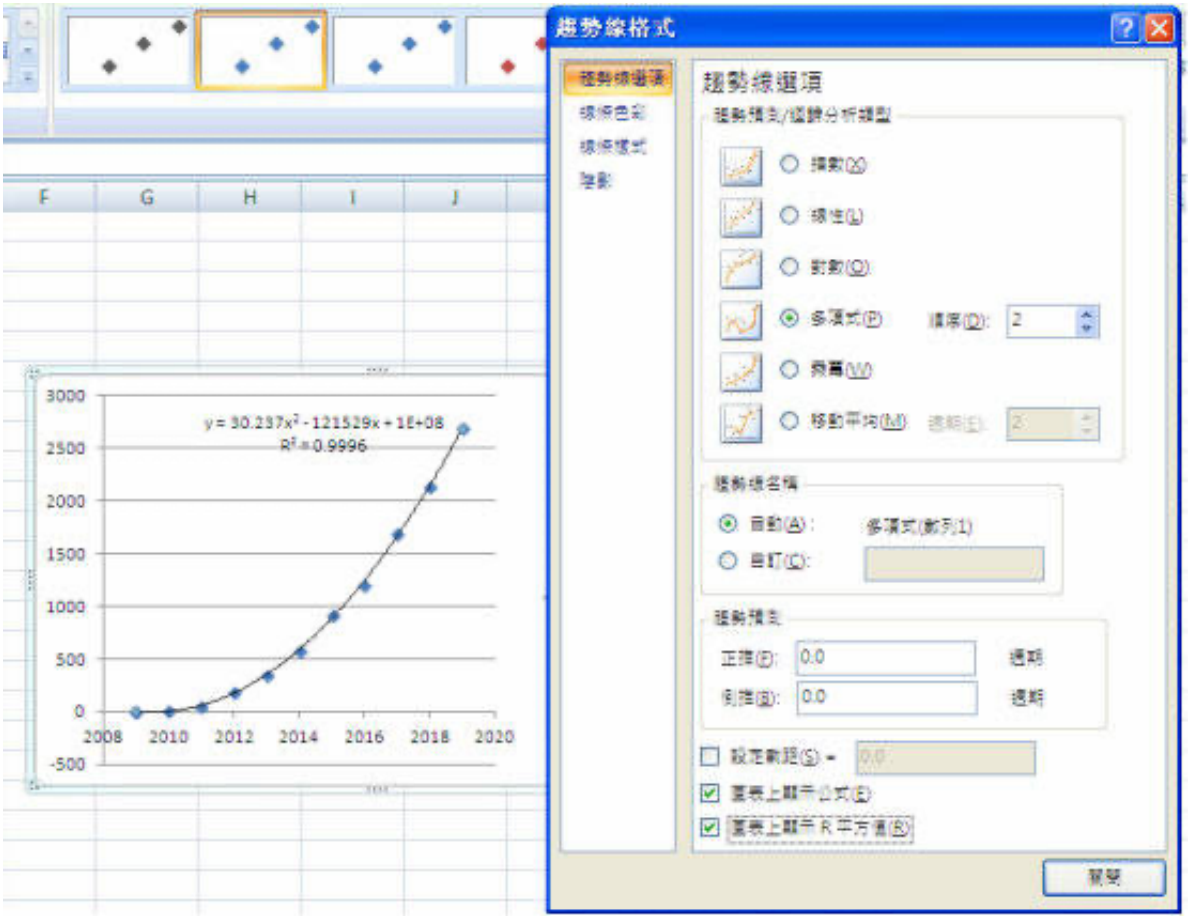
预测 2019 年淘宝双 11 当天销售额：2689.00 亿

预测 2020 年淘宝双 11 当天销售额：3301.51 亿

预测 2021 年淘宝双 11 当天销售额：3980.35 亿

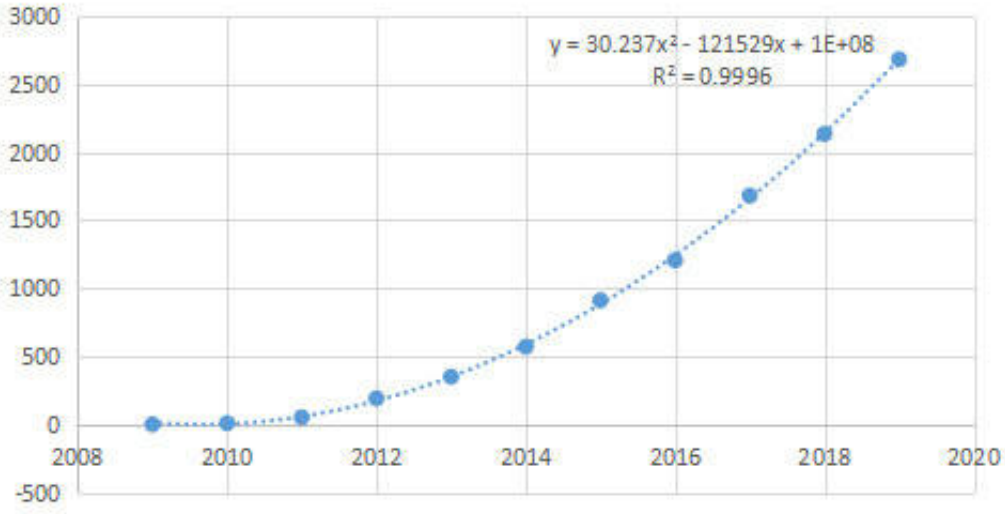
以上这几个图都是引自尹立庆发的文章。我们要解释一下，预测里的二次、三次多项式拟合，以及R-squared是什么意思。不用急着去了解概念，看下面的操作流程自然就明白了。

我自己用Excel可以复制这个二次拟合，截图如下：



这个就是将年份与成交额两列数据，做成一个散点图表。然后鼠标点在一个数据点上，就会出来一个“趋势线”的选项。再把趋势线选择成“多项式”，选2次多项式。再让图表上显示公式、R平方值，左边的曲线拟合图就自动出来了。

其实用国产免费软件WPS里面的表格，也一样可以做出这种趋势线的方程。为支持国产软件，我们用WPS来做。不难摸索出用WPS如何生成趋势线和方程。



上图是WPS生成的2009–2019年11年的成交额数据的拟合曲线，可以看出，WPS和Excel生成的二次拟合方程参数是一样的。

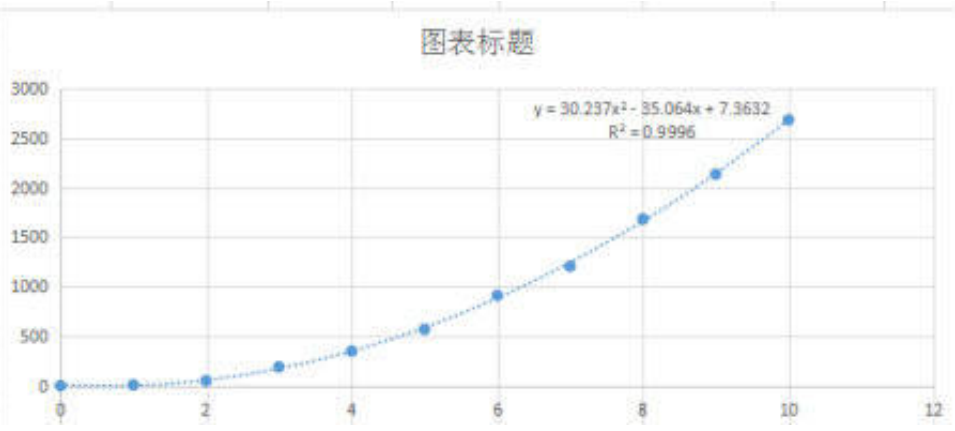
得出的拟合方程是一个二次的多项式:

y = 30.237xx – 121529x + 1E+08，
R平方=0.9996

这个R平方就是“相关系数”，越接近1越好，有一个公式来计算的，后面会解释。公式里的1E+08是科学计数法写的常数项，数值太大了写不下，是一个数字。看样子曲线拟合得很好，但是为什么常数项都大到出不来了？哈哈，因为这个傻软件，把2009–2019当做数值，也就是说x的取值是2009到2019。它不知道是年份，也把这个拟合做出来了。

我们弄聪明点，把年份用0–10代表，2009就是第0年，2019就是第10年，同样把图表和拟合方程做出来。用0开始的好处，是可以直接得到拟合的初值，其实用1–11也差不多。

0	0.5
1	9.36
2	52
3	191
4	350
5	571
6	912
7	1207
8	1682
9	2135
10	2684



这个方程就好多了：

y = 30.237xx – 35.064x + 7.3632
R平方=0.9996

我们把x = 10代进去算，得到的是：
30.237 * 10 * 1 – 35.064 * 10 + 7.3632 = 2680.423

这个数值相当接近2019年天猫的实际成交额2684亿。画在图上这么点差距根本看不出来，点的中心就在趋势线上。看上去拟合得非常好，简直太漂亮了，天猫这11年怎么可能成交得这么准呢？

到此我们可以看出来，所谓的“二次多项式拟合”，就是用一个方程：

Y = A * xx + B * x + C

去拟合一系列x值对应的原始y值，误差越小越好，“拟合度”越接近1越好。这个拟合度，就是用“R平方”来代表的。

我们再把R平方的定义解释一下：

R平方 = 1 – SSE/SST

SSE就是和方差，每个点的拟合值与实际值有一个误差，对它平方，所有点的误差平方加起来，就是SSE。然后所有点原始y值，和平均值有一个差值，对这个差值平方，所有点的差值平方相加，就得到了SST，是个挺大的数。看不懂没关系，我们用下面的表格来解释。

年份	实际成交值	拟合值	误差	误差率(%)
0	0.5	7.3632	-6.8632	-1372.64
1	9.36	2.5362	6.8238	72.903846
2	52	58.1832	-6.1832	-11.89077
3	191	174.3042	16.6958	8.7412565
4	350	350.8992	-0.8992	-0.256914
5	571	587.9682	-16.9682	-2.971664
6	912	885.5112	26.4888	2.9044737
7	1207	1243.5282	-36.5282	-3.026363
8	1682	1662.0192	19.9808	1.1879191
9	2135	2140.9842	-5.9842	-0.28029
10	2684	2680.4232	3.5768	0.1332638
总和	9793.86			
平均	890.3509091			

第一列年份0–10就是公式里的x值，分别对应2009–2019年。第二列就是实际的y值，是每年新闻报出的天猫成交额。这些实际成交值，有一个平均值890.35。拟合值，就是用公式“**y = 30.237xx – 35.064x + 7.3632**”算出来的每一年的值。误差，就是用“实际成交值”减去拟合的值。

误差平方	与均值差异	均值差异平方
47.10351	-889.851	791834.8
46.56425	-880.991	776145.14
38.23196	-838.351	702832.4
278.7497	-699.351	489091.82
0.808561	-540.351	291979.2
287.9198	-319.351	101985.06
701.6565	21.649	468.6792
1334.309	316.649	100266.59
399.2324	791.649	626708.14
35.81065	1244.649	1549151.1
12.7935	1793.649	3217176.7
3183.18		8647639.7
SSE		SST
0.999632		

这个表是与上个图并排的，分开来看得清楚一些。误差平方就是对前面得到的误差值进行平方。所有的误差平方相加，就是误差平方和SSE，等于3183.18。

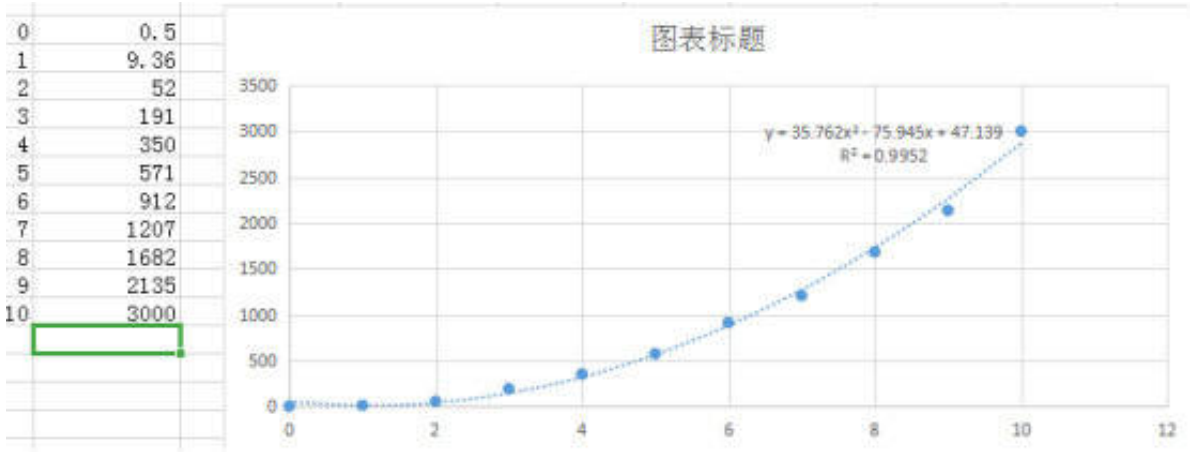
每一年的实际成交值，和平均值890.35求出一个差异。然后对每个值平方，再求和。得到了“均值差平方和”SST，数值很大是8647639.7。

然后就可以得到“相关系数”R平方，是 1 – 3183.18/8647639.7 = 0.9996。这个值就正好和WPS生成的R平方值相等。

看这个表格，我们就忽然发现，**这个拟合的“神奇感”好象下降了**。你看第一年0.5的值，拟合值7.3632却是它的十多倍。第二年的拟合值2.5362，甚至不到第一年拟合值的一半。如果第二年业绩是这个鬼样，马云得把天猫负责人就地撤职。但是画成图，因为绝对坐标的关系，早期很大的拟合误差，与以后很大的成交值相比显得很小，画出来显得拟合得很好。

我们发现，越是早期的小数据，误差相对越大。定义一个误差率，是误差值除以原值乘以100%。头四年的误差离谱地大，高的有1372%和72.9%，第四年差异仍然高达8.74%。后面随着原值的逐渐地大，误差率就逐渐减小，只有0.2%、0.1%了。

也就是说，这个拟合的“秘诀”是：**注意把每个点的误差的绝对值弄得小一些，顾头不顾腚，顾大不顾小**。最后画出图来，因为坐标要跟比较大的数值的尺度，前面较大的相对误差就缩起来看不见了。



另一点要注意的，**不要迷信那个R平方值**，以为多么接近1啊，真神奇啊。例如上图，假设2019年天猫成交额不是2684亿，而是跑到一个很歪的3000亿去了。那么我们新做一个二次多项式拟合，**得出的R平方值仍然有0.9952**，还是相当接近1。从图上看出来，后面两个点已经有点偏了，R平方值仍然漂亮得很。这是因为这个R平方值，分母SST是个特别大的数，怎么算最后总是接近于1。

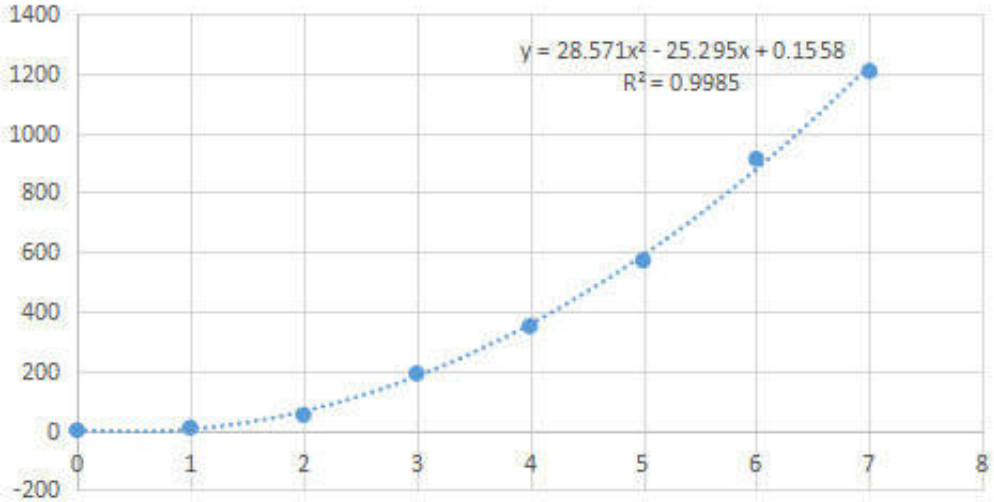
尹立庆的预测巧合在于，他对2019年的预测正好碰上这年天猫的增长是中规中矩的25.7%。二次多项式拟合的预测值2675，预期增长是25.3%，正好相差不大。天猫2019年成交增长25%，这个并不奇怪，不少人随口说个直觉也可能是这个数。

但并不是每一年都如此，其实天猫的增长率也出过异常。

年份	成交额	增长率（%）
2009	0.5	
2010	9.36	1772
2011	52	455.5555556
2012	191	267.3076923
2013	350	83.2460733
2014	571	63.14285714
2015	912	59.71978984
2016	1207	32.34649123
2017	1682	39.35376968
2018	2135	26.93222354
2019	2684	25.71428571

看上图天猫历年的增长率，前面增长率高，后面增长率逐渐下滑，因为规模大了增长率下跌正常。但是2017年增长了39%，高于2016年的32%，这是一个数据异常。

假设我们在2016年，看到8年的成交数据，搞了一个二次拟合，结果会是如何？



我们用8个点，同样得到了一个相当漂亮的拟合曲线！R平方值也是0.9985，相当接近于1。方程是：

y = 28.571xx – 25.295x + 0.1558

如果用这个二次多项式方程，去算2017年的值，会是：

y = 28.571*8*8 – 25.295*8 + 0.1558 = 1626.34

2017年的实际成交额是1682亿，差了50多亿，就没有2019年只差几亿那么神了。

有趣的是，这个公式对2017年的预测增长率是34.74%，也高于上年的32.35%。这是因为，2014、2015、2016三年的增长率分别是63%、59%、32%。这个32%降得有点多，在下一年就补回来一些。比如天猫管理团队认为，2016年增长率不尽如人意，要多想招，2017年的增长率就搞到了39.35%，发力过度，比拟合预测的还要高了。

让我们来看8个点和11点得到的两个二次多项式拟合方程：

2009–2016： y = 28.571xx – 25.295x + 0.1558
2009–2019： y = 30.237xx – 35.064x + 7.3632

注意，这两个方程对应的三个系数，差异已经非常大了。就算马云有一个“按公式操纵天猫每年双11成交额”的邪恶计划，我们也搞不清楚他最初设计的二次方程系数是如何的。

所以，要么马云没有操纵天猫双11成交额的数学方程，要么马云在动态修正预测成交的方程。不太可能在某年就把这些系数定死了。

其实，**马云在动态修正预测成交的方程**，这个倒是接近真相了。本来做生意就是这样的，**上一年增长够高了，下一年的增长任务就轻一点，以免各种配套跟不上；上一年增长觉得低了，下一年就多努力做高，免得业绩增长不好看**。但是都动态上了，本就无可厚非，是人家在搞数值化管理，谁管得着？

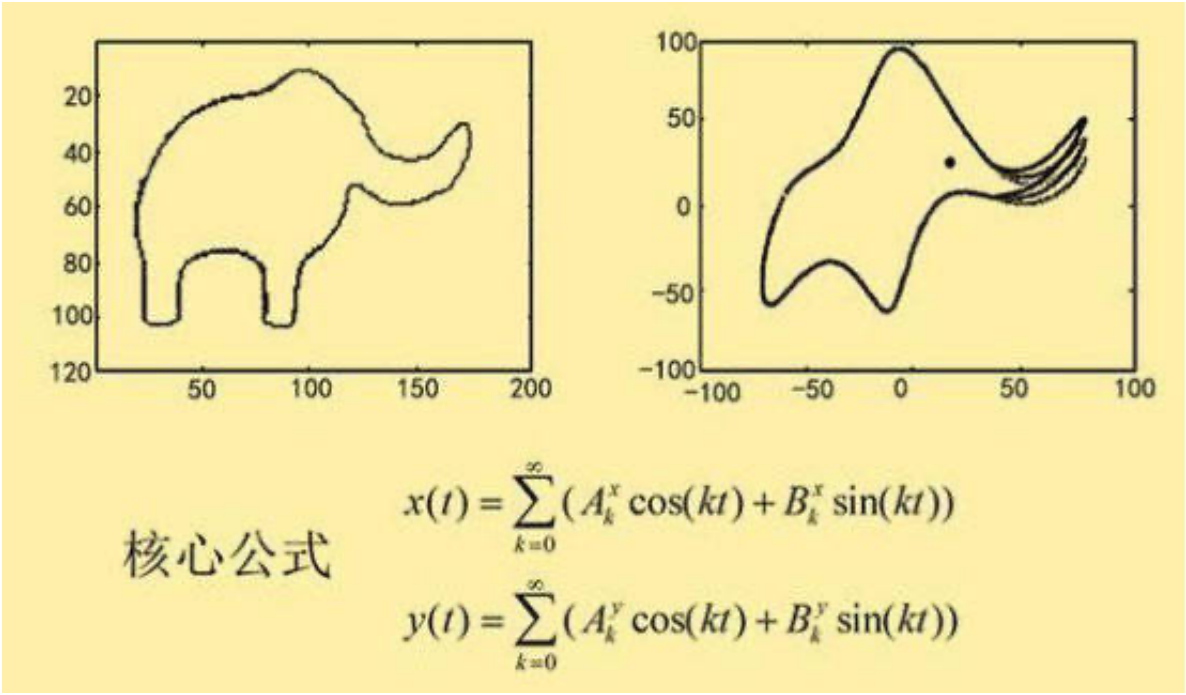
等数值都出来了，再去回头把二次多项式的系数拟合出来，我们可以发现，很容易就拟合得不错，而且R平方相关系数可以做得很漂亮。前提条件是，这一系列数据增长率要比较大，前期的数据比较小，后期的数值大，就可以仅用二次多项式做出一个漂亮曲线了。

如果增长率变动有点大，那就要用三次、四次多项式了。但原理是一样的，就不再分析了。

总之，不要相信拟合的神奇，也不要相信“拟合度”接近1的神奇效果。这在数学统计里，实在很平常。对搞过数据分析的人来说，这是最平常的手段，迷信拟合真是少见多怪，只会被内行笑掉大牙。

这还只是二次多项式三个系数的拟合。要是用深度学习那上百万个系数来拟合，结果可以

漂亮得让一些传统研发人员怀疑人生，转而去搞机器学习。



最后，贴个经典拟合搞笑图：大象。给吃瓜群众看看，数学公式拟合的威力有多大。