评委一评分，签名及备注	队号： 1010	评委三评分，签名及备注
评委二评分，签名及备注	选题： B	评委四评分，签名及备注
题目：	公司业务数据分析	

摘要

公司业务的发展是通过各个指标来具体体现的，本篇论文主要通过统计学的知识对提供的数据进行合理的计算处理，得出相应的结论

模型一 多元线性回归模型 首先我们可以利用散点图来对 5 个业务进行大致的判断，根据曲线趋势判断是否达到饱和状态，这样可以大大简化运算，不过这相对来说存在一定的误差，所以下面通过建立精确的回归方程进行拟合分析预测未来发展是否会达到饱和并求出具体的饱和值

模型二 相关系数模型 根据模型一中各项指标建立的散点图来判断出哪项指标是收入指标。公司追求经济效益自然要判断出哪项业务有促进作用了，我们可以建立相关系数模型，通过判断收入与业务的相关系数的值的大小来得出结论，相关系数值越大则两者之间的相关性越高

模型三 正交设计模型 各项业务是不可能孤立存在的，公司要想发展壮大必须会紧密联系各项业务时期发挥出最大的功效，我们可以据此建立正交设计模型，得出正交实验值判断出哪几项业务可以相互促进使得收入增加

以上三个模型的建立是在公司提供的数据是连续的情况下的出来的，但是这和实际情况不符，故在模型的改进与推广中建立偏最小二乘 log*istic* 回归模型进行深化运算得出更加准确的结论来

最后在以上四个模型得出结论的基础上，分析当前公司的状态并就如何扩大公司的盈利空间级服务规模提供实质性的建议，

关键字： 多元线性回归模型、相关系数模型、正交设计模型、偏最小二乘 log*istic* 回归模型、统计学

公司业务数据分析

一、问题重述

某互联网公司推出一项服务，此项服务包括 5 个主要的业务，这 5 项业务共包含 8 个指标，某项业务可以含有 1 个或多个指标，在这 8 个指标中其中有一个指标是收入。客户可以根据自己的需要选择开通某些业务，各个业务之间没有强制绑定关系，但是某些业务之间通过相互宣传有一定的促进作用。附件中是本公司 2012 年第一季度的数据，包括各个业务的各个指标的数据：指标数据为 0，说明该业务还没有这个指标；从 0 变为正数说明此项业务开始包含新的功能，新功能具有新的指标。附件中还包括此项服务带来的收入数据。

请你根据各个服务的指标数据和收入数据，完成如下问题：

1、其中某些业务的使用量接近饱和，请你建立模型计算哪些业务量接近饱和，饱和的指标估计值是多少；

2、根据财务数据，你能判断出哪个指标是收入吗，请你说明收入主要和哪些业务相关；

3、请你分析出各个业务之间的相关性，哪几个业务相互促进可以使得收入增加；

4、假如你是本服务的项目经理，根据现有的数据和你所建立的模型，给公司总经理写一份季度分析报告，分析当前的状态以及以后发展的建议，如何扩大公司的盈利空间以及服务规模。

二、问题分析

题目最终的宗旨就是希望我们解决如下三个题目后能够在第四个题目中给出实质性的建议，即该公司在接下来的时间内如何发展，如何扩大企业规模，如何弥补业务的缺陷从而能够带来最大的经济效益

1.问题 1 希望我们能够发现哪些业务相对来说不受到顾客欢迎带来的经济效益较小从而公司管理者可以在接下来的时间内进行弥补改进。我们可以利用散点图大致的进行判断，不过这相对来说存在一定的误差，所以仍需建立精确的回归方程进行拟合分析预测未来发展是否会达到饱和

2.问题 2 要求我们能够判断出哪项指标为收入即可但要求合理有据。现实情况下公司的收入应始终是处于上升的趋势当然可以有小的波动说明企业此时处于亏损但不能是长时间的，那么我

3.各项业务是不可能孤立存在的，公司要想发展壮大必须会紧密联系各项业务时期发挥出最大的功效，我们可以据此建立正交设计模型，得出正交实验值判断出哪几项业务可以相互促进使得收入增加

4.问题 4 实际上是要求我们在顺利解决了上述三个题目后能够给公司提出一些实质性的建议利于公司以后的进一步大发展。

二、模型假设

1. 在数据计算过程中，在误差在合理范围内，对结果影响较大的数据可以忽略不讨论；
2. 当统计数据得到的结果与实际出入较大时，可以忽略该因素的影响而暂不考虑；
3. 假设公司处于正常运作状态并为遇到金融危机之类的事件

四、符号说明

β _____回归系数

$\hat{y}$ _____回归值

Q _____偏差平方和

$\bar{x}_j$ _____样本均值

S_j _____样本标准差

E_x _____期望值

V _____协方差矩阵

var(y) _____方差

R^2 ••••••••　判定系数

SSR ••••••••　回归平方和

SST ••••••••　残差平方和

五、模型建立

5.1 模型一：多元回归模型

5.11 建立多元线性回归模型，总体回归函数的一般形式如下

$$Y_t = \beta_1 + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + u_t$$

上式假定因变量 Y 与 $(k-1)$ 个自变量之间的回归关系可以用线性函数来近似反应。式中：Y_t 是变量 Y 的第 t 个观测值；X_{jt} 是第 j 个自变量 X_j 的第 t 个观测值 $(j=2,\ldots,k)$；u_t 是随机误差项；$\beta_1,\beta_2,\ldots,\beta_k$ 是总体回归系数。β_j 表示在其他自变量保持不变的情况下，自变量 X_j 变动一个单位所引起的因变量 Y 平均变动的数额，因而又叫偏回归系数。该式中，总体回归系数是未知的，必须利用有关的样本观测值来进行估计。

假设已给出了 n 个观测值，同时 $\hat{\beta}_1,\hat{\beta}_2,\ldots,\hat{\beta}_k$ 为总体回归系数的估计，则多元线性回归模型的样本回归函数如下 $Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \ldots + \hat{\beta}_k X_{kt} + e_t \ (t=1,2,\ldots,n)$ 式中：e_t 是 Y_t 与其估计 Y_t 之间的离差，即残差。

5.12 显著性检验

[1] 回归系数的显著性检验

多元回归中进行这一检验的目的主要是为了检验与各回归系数对应的自变量对因变量的影响是否显著，以便对自变量的取舍做出正确的判断。一般来说，当发现某个自变量的影响不显著时，应将其从模型中删除。这样才能够做到以尽可能少的自变量去达到尽可能高的拟合优度。

多元模型中回归系数的检验同样采用 $t-$ 检验和 $P-$ 检验，其原理和基本步骤与一元回归模型基本相同。下面仅给出回归系数显著性检验 t 统计量的一般计算公式 $t_{\hat{\beta}_j} = \dfrac{\hat{\beta}_j}{S_{\hat{\beta}_j}} (j=1,2,\ldots,k) \quad (1)$

式中：$\hat{\beta}_j$ 是回归系数的估计值；$S_{\hat{\beta}_j}$ 是 $\hat{\beta}_j$ 的标准差的估计值。$S_{\hat{\beta}_j}$ 按下式计算

$$S_{\hat{\beta}_j} = \sqrt{S^2 \times \psi_{jj}}$$

式中：ψ_{jj} 是 $(X'X)^{-1}$ 的第 j 个对角线元素；S^2 是随机误差项方差的估计值。(1) 式的 t 统计量背后的原假设是 $H_0: \beta_j = 0$，因此 t 的绝对值越大表明 β_j 为 0 的可能性越小，即表明相应的自变量对因变量的影响是显著的。

[2]回归方程的显著性试验

多元线性回归模型包含了多个回归系数，因此对于多元回归模型，除了要对单个回归系数进行显著性检验外，还要对整个回归模型进行显著性检验。由离差平方和的分解公式可知，回归模型的总离差平方和等于回归平方和与残差平方和的总和。回归模型总体函数的线性关系是否显著，其实质就是判断回归平方和与残差平方和之比值的大小问题。由于回归平方和与残差平和的数值会随观测值的样本容量和自变量个数的不同而变化，因此不宜直接比较，而必须在方差分析的基础上利用 F 检验进行。其具体的方法步骤可归纳如下。

（1） 假设总体回归方程不显著，即有 $H_0=\beta_2=\beta_3=\ldots=\beta_k=0$

（2） 进行方差分析，列出回归方差分析表

离差名称	平方和	自由度	方差
回归平方和	$SSR=\sum\left(\hat{Y}_t-\bar{Y}\right)^2$	$k-1$	$SSR/(k-1)$
残差平方和	$SSE=\sum e_t^{\,2}$	$n-k$	$SSE/(n-k)$
总离差平方和	$SST=\sum\left(Y_t-\bar{Y}\right)^2$	$n-1$	

表中，回归平方和的取值受 k 个回归系数估计值的影响，同时又要服从 $\sum\hat{Y}_t/n=\bar{Y}$ 的约束条件，因此其自由度是 $k-1$。残差平方和取决于 n 各因变量的观测值，同时又要服从 k 个正规方程式的约束，因此其自由度是 $n-k$。回归平方和与残差平方和各除以自身的自由度得到的是样本方差。

（3） 根据方差分析的结果求 F 统计量，即 $F=\dfrac{SSR/(k-1)}{SSE/(n-k)}$

数学上可以证明，在随机误差项服从正态分布同时原假设成立的条件下，F 服从于自由度为 $(k-1)$ 和 $(n-k)$ 的 $F-$分布。

（4） 根据自由度和给定的显著性水平 α，查 $F-$分布表中的理论临界值 F_α。当 $F>F_\alpha$ 时，拒绝原假设，即认为总体回归函数中各自变量与因变量的线性回归关系显著。当 $F<F_\alpha$ 时，接受原假设，即认为总体回归函数中，自变量与因变量的线性关系不显著，因而所建立的模型没有意义。

5.13 多元线性回归预测

在通过各种检验的基础上，多元线性回归模型可以用于预测。多元线性回归

预测与一元线性回归预测的原理是一致的，其基本公式如下

$$\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_{2f} + \ldots + \hat{\beta}_k X_{kf}$$

式中：$X_{jf}\ (j = 2,3,\ldots,k)$是给定的$X_j$在预测期的具体数值；$\hat{\beta}_j$是已估计出的样本回归系数；$\hat{Y}_f$是$X_j$给定时$Y$的预测值。

5.2 模型二：相关系数模型

相关系数（Correlation Coefficient）:度量变量间相关关系的一类指标的统称。就参数统计而言，常用的是皮尔逊积矩相关系数（Pearson）:即协方差与两变量标准差之间的比值，是没有量纲的，标准化的协方差。
协方差（Covariance）：两个变量与其均值离差乘积的平均数，是相互关系的一种度量。

总体协方差：$\sigma_{xy} = \dfrac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$

样本协方差：$S_{xy} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

根据样本相关系数的计算公式有：

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x\text{^}2 - (\sum x)\text{^}2} \bullet \sqrt{n\sum y\text{^}2 - (\sum y)\text{^}2}}$$

相关系数的特点
样本相关系数r有以下特点。

（1） r的取值介于-1与1之间。

（2） 当$r=0$时，X与Y的样本观测值之间没有线性关系

（3） 在大多数情况下，$0<|r|<1$，即X与Y的样本观测值之间存在着一定的线性关系。当$r>0$时，X与Y为正相关。当$r<0$时，X与Y为负相关。

（4） 如果$|r|=1$，则表明X与Y完全线性相关。当$r=1$时，称为完全正相关；而$r=-1$时，称为完全负相关。

（5） r是对变量之间线性相关关系的度量。$r=0$只是表明两个变量之间不存在线性关系，它并不意味着X与Y之间不存在其他类型的关系。对于两者之间可能存在的非线性相关关系，需要利用其它指标去进行分析。

5.3 模型三 ： 正交设计模型

双因子方差分析模型。在这两种模型中，试验数据的统计分析有以下两大优点：

1） 因子水平（或水平组合）参数的估计有简单的表达形式；

2） 因子效应（包括主效应和交互效应）和随机误差效应可以用平方和分解的方法进行分离，进而用 F 统计量进行检验。

在此我们要指出两种模型的一个重要区别：对单因子方差分析模型，我们不要求在每个水平上的试验次数相同；而对双因子方差分析模型，在每对因子水平组合上，试验的平衡性（即等重复性）是一个重要条件，不然的话，平方和分解公式就不成立，这样在方差分析时就会产生一定的困难。在多因子试验中也有同样的问题。因此，我们只考虑平衡的多因子试验。

双因子试验的方差分析模型中所包含的统计思想和方法可以一般地推广到多因子试验的场合。以三因子模型为例，设有三个因子对响应变量有影响，分别记为 A、B、C，它们的水平数分别为 I、J、K。全面地考虑，这三个因子对响应变量的影响可以分成以下三种：

1） 各因子的主效应，即单个因子的不同水平对响应变量产生的影响；

2） 一阶交互效应（双因子交互效应），即在扣除主效应的影响之后，任意两个因子的不同水平组合（AB、AC、BC）对响应变量产生的联合影响；

3） 二阶交互效应（三因子交互效应），即在扣除主效应和一阶交互效应的影响之后，三个因子的不同水平组合（ABC）对响应变量产生的联合影响。与双因子的情况类似，如果在三个因子的每个水平组合上作相同的 L 次试验，则当 L>1（有重复）时，可以用全模型（即包含全部上述三种效应的模型）进行方差分析；而当 L=1（无重复）时，二阶交互效应无法分析，而只能分析主效应和一阶交互效应。读者可以仿照上一节中的作法，对这两种情况下三个因子方差分析的全部过程列出结果（模型、平方和分解、自由度、F 统计量，等等）。进而可以考虑四因子、五因子、乃至一般 m 个因子的情况。无论有多少个因子，如果在所有因子的每个水平组合上都作至少一次试验，则试验是完全的。为便于进行方差分析，试验应该是等重复的。为能够够分析最高阶（*m-1* 阶）交互效应，试验应该是有重复的（重复数大于 1）。

虽然我们在理论上可以容易地将双因子方差分析的模型和方法推广到多因子方差分析的情况，但是，在实践中，作多个因子的完全试验会有实际的困难，因为完全试验所要求的试验次数太多，乃至无法实现。例如，假定要考虑五个三水平因子，则完全试验（重复数为1）要求作35=243 次试验；假如再加一个四水平因子，则完全试验（同样重复数为1）要作972 次试验。如果要能够分析全部交互效应，同时还能够作平方和分解，则试验次数还需加倍！显然，如此大的试验次数在实际中几乎是无法实施的。如何解决这个困难呢？我们先提出如下的思路供思考。

在对一个因子试验所建立的线性模型中，独立参数（总均值、主效应、交互效应等）的个数 k 与试验次数 n 之间有下面的关系：当 n>k 时，有足够的自由度 k 来估计参数，同时还有剩余自由度来估计误差的方差（n-k>0）；当 n=k 时，有足够的自由度来估计参数，但是没有剩余自由度来估计误差的方差 n-k=0；当 n<k 时，没有足够的自由度来估计参数，同时也没有自由度来估计误差的方差。在因子试验中，除非可以事先确定数据中的随机误差很小，以至可以简单地忽略，否则误差的估计是必要的，它是进行 F 检验的前提。因此，如果不能简单忽略随机误差，就应该给误差的估计留下适当的自由度（n>k）。对这样一个思路，我们不想在此作理论上的论证，读者可以结合双因子试验中有重复和无重复的两种情况来领会。在双因子有重复试验中，试验次数大于交互效应模型中独立参数

的总数，因此有剩余的自由度来估计误差方差；而在双因子无重复试验中，试验次数等于交互效应模型中独立参数的总数，因此没有剩余自由度来估计误差方差。此时，要估计误差就只能用可加效应模型。

根据上述的思路，只要试验总次数N大于独立参数的个数M就可以有足够的自由度来估计参数，同时还有剩余的自由度来估计误差方差，进而作假设检验。这是因子试验设计中要考虑的第一件事。第二件事是要使参数估计和检验统计量有好的性质和形式，关键是要使各组效应的参数估计之间相互独立，同时使相应的平方和之间相互独立。但是，在一个线性模型中，参数（主效应及各种交互效应）的数目是由实际问题本身决定的，而不是由人主观决定的。在大量的因子试验的实践中，人们发现：在很多情况下，因子之间只有主效应，至多存在某些一阶交互效应（即两因子的交互效应）。高阶交互效应在很多情况下是不存在的。在这种情况下，多因子试验的模型中包含的参数实际上并不多，可能远远少于全模型的参数。比如有 6 个二水平因子，如果考虑所有可能的交互作用就有 26=64 个独立参数（包括总均值），但是如果只考虑主效应则只有 6+1=7 个立参数。因此对 6 个二水平因子的可加效应模型，理论上只需作 8 次试验就可以有多余的自由度来估计误差方差。

如何安排试验，使得上述的两个想法很好地实现呢？从双因子无重复试验的可加模型的分析中得到启示。在这个模型中，由于两个因子的所有水平组合都作了相同次试验（一次），因此两组因子主效应的参数估计不仅有简单的形式，而且还是相互独立的，因而平方和之间也是相互独立的。因此，对于多因子试验的无交互效应模型（只考虑主效应），如果我们能如此安排试验，使得对任何一对因子，它们的所有水平组合都作了相同次试验，则对任何一对因子，两组因子主效应的参数估计和平方和也应具有上述性质。进而，如果试验的总次数 n 超过参数的总个数 k，则还有多余的自由度来估计误差，进行方差分析。实际上，这就是“正交因子设计”原理的基本思路。

六、模型求解

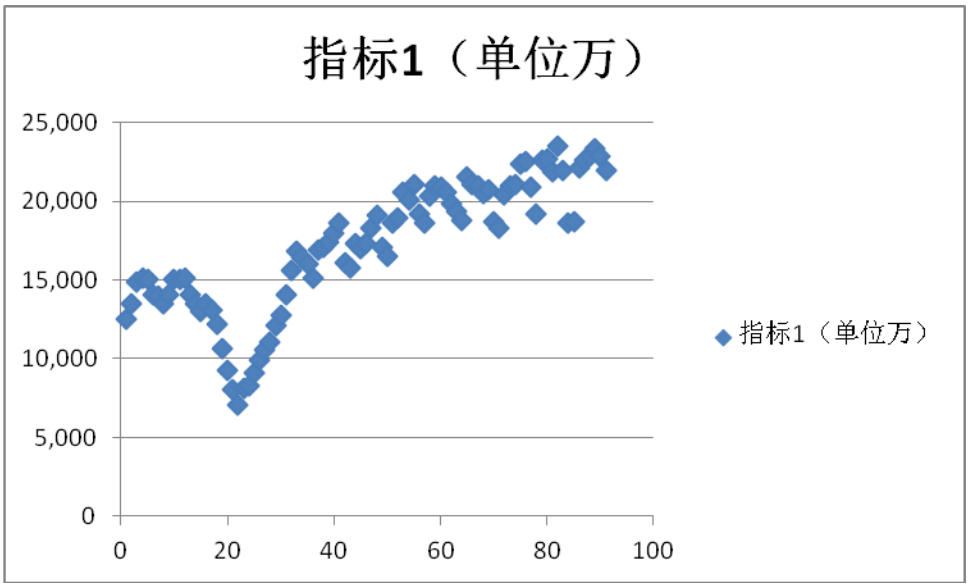
6.1 模型一的求解

题目要求判断出哪些业务的使用量接近饱和，可以在现有 5 个业务数据的基础上建立多元线性规划模型，根据散点图数据的发展趋势，大致的判断出部分业务是否出现饱和状态，这样可以简化运算

下面对各个业务进行分析判别是否饱和

6.11 业务 1

根据业务 1 的数据画出三点图，数据见附录 1



图一　业务 1 散点图

通过散点图不难看出指标 1 始终在处于上升状态，只是随着时间的增长增长趋势有所延缓，但并未达到饱和状态，为了进一步验证自己的想法，下面我们建立回归模型进行数值检验。

表 1 业务 1 回归估计结果

回归统计	
Multiple R	0.838788529
R Square	0.703566197
Adjusted R Square	0.70023548
标准误差	14.46152529
观测值	91

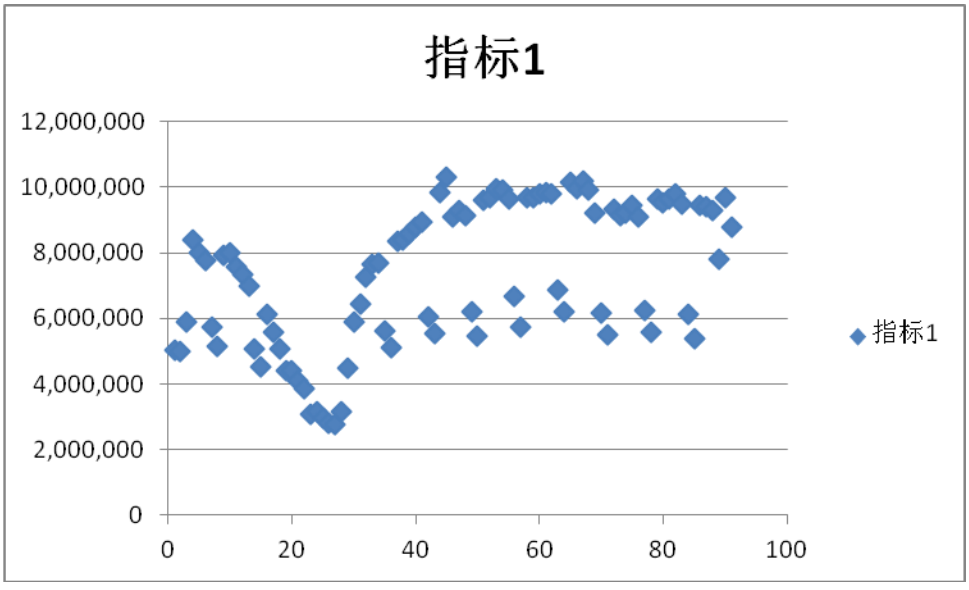
方差分析				
	df	SS	MS	F
回归分析	1	44176.92	44176.92	211.2357
残差	89	18613.08	209.1357	
总计	90	62790		

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-45.1058	6.449192	-6.99403	4.76E-10	-57.9202	-32.2914
X Variable 1	0.005322	0.000366	14.53395	3.17E-25	0.004595	0.00605

根据以上结果，可得业务 1 的样本回归方程为 $Y=-45.1058+0.005322X$

F 的统计量为 211.2357，说明模型整体的线性关系非常显著。参数值与理论分析的结论相同，同时 $P-$值均很小，因此在 1%的显著性水平上，这几个参数均能通过显著性检验。由模型的回归方程可知，该函数为单调递增函数自然不存在饱和之说了

项目二



图二　业务 2 散点图

同理根据业务 1 的判断准则得出业务 2 的回归估计结果

表 2 业务 2 的回归估计结果

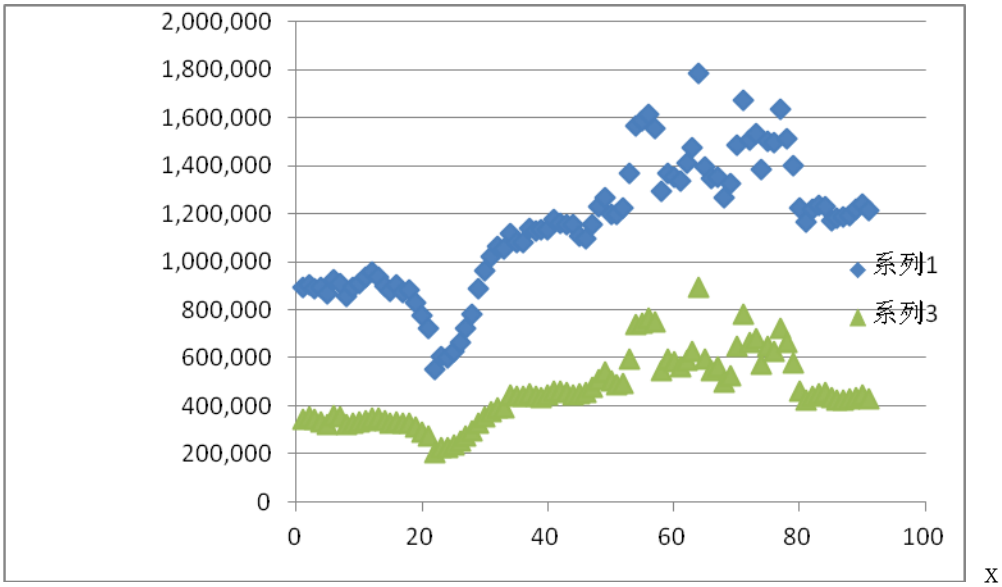
回归统计	
Multiple R	0.82237
R Square	0.676292
Adjusted R Square	0.668935
标准误差	15.1978
观测值	91

方差分析					
	df	SS	MS	F	Significance F
回归分析	2	42464.37	21232.19	91.92495	2.8E-22
残差	88	20325.63	230.9731		
总计	90	62790			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-94.9176	10.68133	-8.8863	7.04E-14	-116.144	-73.6907
X Variable 1	-2.4E-05	3.04E-06	-7.85639	9.1E-12	-3E-05	-1.8E-05
X Variable 2	0.00021	2.04E-05	10.27457	9.84E-17	0.000169	0.00025

根据以上结果，可得业务 2 的回归方程为 $Y=-94.9176-(2.4\mathrm{E}-05)X_1+0.00021X_2$. 不难发现两个指标的系数相同，则最终会出现饱和状态，代入数值计算得出饱和值为-1184

项目三



图三 业务 3 散点图

表 3 业务 3 的回归估计结果

回归统计	
Multiple R	0.849462
R Square	0.721586
Adjusted R Square	0.711985
标准误差	14.17527
观测值	91

方差分析					
	df	SS	MS	F	Significance F
回归分析	3	45308.38	15102.79	75.16139	4.47E-24
残差	87	17481.62	200.9382		
总计	90	62790			

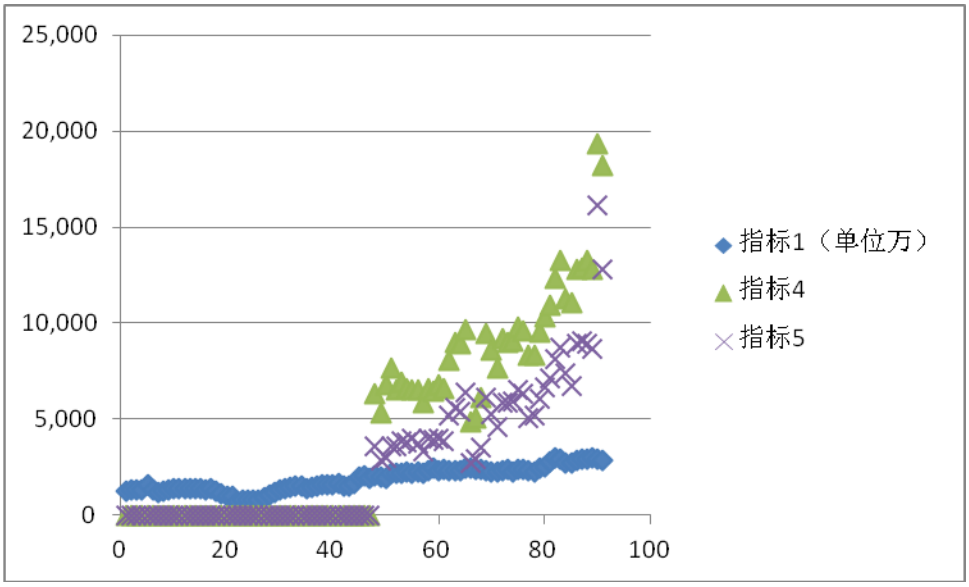
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-78.7024	8.810041	-8.93326	6.15E-14	-96.2133	-61.1915
X Variable 1	0.000154	6.01E-05	2.57114	0.011835	3.51E-05	0.000274
X Variable 2	0.000206	0.000129	1.599437	0.113349	-5E-05	0.000461
X Variable 3	-0.00046	9.58E-05	-4.83657	5.66E-06	-0.00065	-0.00027

根据以上结果，可得业务 3 的回归方程为

$$Y = -78.7024 + 0.000154X_1 + 0.000206X_2 + -0.00046X_3$$

不难发现三个系数和相加接近为零，说明最终会出现饱和状态，代入数值算出饱和值为 117.8

项目四

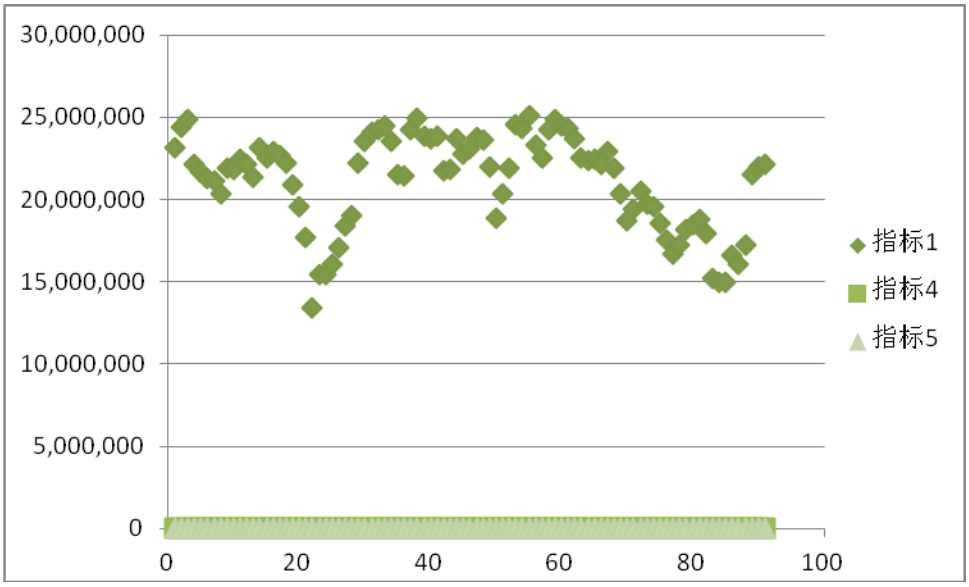


，，

图四　业务 4 散点图

由散点图可知业务 4 始终处于上升趋势，不可能出现饱和状态

项目 5



图五　业务 5 散点图

表 4 业务 5 的回归估计结果

回归统计	
Multiple R	0.463315
R Square	0.214661
Adjusted R Square	0.18758
标准误差	23.80752
观测值	91

方差分析					
	df	SS	MS	F	Significance F
回归分析	3	13478.56	4492.852	7.926723	9.84E-05
残差	87	49311.44	566.7982		
总计	90	62790			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	109.6585	19.26177	5.693066	1.66E-07	71.37367	147.9434
X Variable 1	-4.1E-06	9.66E-07	-4.28169	4.76E-05	-6.1E-06	-2.2E-06
X Variable 2	0.001114	0.000361	3.086016	0.00272	0.000397	0.001832
X Variable 3	-0.00101	0.001958	-0.51646	0.606844	-0.0049	0.00288

根据系数值大小进行判断出业务 5 也不可能达到饱和状态

6.2 模型二的求解

6.21 判断收入

由模型一作出的散点图和回归方程很容易看出业务 1 中的指标即为收入，因为该数值始终处于上升状态且数值较大满足公司的运作要求

6.22 判断收入与那些业务相关

将收入数值单独调出来分别与各个业务进行相关系数运算，根据各个系数值的大小判断收入与该业务是否相关

表 5 收入与其他四个业务的相关系数，

收入与项目 2 之间的相关系数

	列 1	列 2	列 3
列 1	1		
列 2	0.824558	1	
列 3	0.917061	0.970894	1

收入与项目 3 之间的相关系数

	列 1	列 2	列 3	列 4
列 1	1			
列 2	0.834725	1		
列 3	0.778724	0.993469	1	
列 4	0.712583	0.975052	0.990786	1

收入与项目 4 之间的相关系数

	列 1	列 2	列 3	列 4	列 5
列 1	1				
列 2	0.942102	1			
列 3	0.955506	0.92119	1		
列 4	0.7996	0.90959	0.736482	1	
列 5	0.774348	0.883385	0.705028	0.989865	1

收入与项目 5 之间的相关系数

	列 1	列 2	列 3	列 4
列 1	1			
列 2	0.09993	1		
列 3	0.299981	0.415745	1	
列 4	0.206405	0.374574	0.728762	1

根据相关系数数值的大小，不难看出收入主要和业务 2、业务 3 和业务 4 相关而与业务 5 无关

6.3 模型三的求解

考虑设计一个试验，安排m 个因子，作n 次试验，若它满足下面两个条件，则这个试验称为正交试验：

1) 每一因子的不同水平在试验中出现相同次数（均衡性）；

2) 任意两因子的不同水平组合在试验中出现相同次数（正交性）.

就定义来说，等重复的完全试验显然满足条件，因此当然是正交试验。但是，如果因子的水平数分别为$t1$, $t2$, … , tm, 则完全试验至少要作$N=t1,t2,\cdots tm$次试验，由于要求的试验次数太多，实际上很难实施。我们通常所说的正交试验设计是指既满足上述两条件，同时试验次数n 又远远小于N 的设计。

正交试验设计的方案可以用一张表来表示, 这张表就称为正交设计表.一般, 正交设计表第一行为表头, 标明每列所代表的因子, 最左一列标明试验的序号(并不表示试验的时间先后顺序, 先后顺序要按照随机化原则来安排), 由1 到n. 表中每列中的数字代表相应因子的水平序号; 每行的数字代表在相应试验中各因子的水平序号. 在正交设计表中：

1) 每列中不同数字出现的次数相同(试验的均衡性);

2) 每两列中不同的数字组合出现的次数相同(试验的正交性).

这两条性质符合正交试验设计的定义.假定因子对响应变量的影响无交互效应（许多实际情况正是这样），正交试验的优点是在很少的试验次数（与全面试验相比）中，所得数据可以简便而有效地对因子效应进行参数估计和方差分析。其方法可一般地归纳如下:

1) 总均值的估计=试验数据的总平均值,

2) 某因子的某个主效应的估计=该因子的该主效应所出现的试验数据的平均值-总平均值,

3) 总平方和=(试验数据-总平均值)的平方和, 自由度=n-1,

4) 某因子的主效应平方和=重复数×参数估计的平方和, 自由度=水平数-1,

5) 残差平方和=总平方和-(因子效应平方和的和), 自由度=总平方和-(因子效应自由度的和).

通过正交表格的计算，我们发现业务 2 和业务 3，业务 2 和业务 4 能够相互促进并使得收入增加，而业务 5 则与其他业务没有多少相互促进

七、结果分析与检验

为了进行合理的模型检验，准确考察模型的正确性和有效性，以 2012 年 3 月 31 日的相关统计数据作为最好的实际状况来验之，根据最后检验的结果与模型求解的结论相比较，发现所建立模型得出的结论可靠性较高。

普通线性回归模型的因变量取值必须是连续的，而公司运营状态是变化的这就无法进行保证，对于此类事件以上所建立的模型不再起作用，原因很简单因为自变量的改变因变量也应随之改变，为此我们需要重新建立函数关系组成一个新模型偏最小二乘$\log istic$回归模型，同样，我们依然可以利用 Excel 进行回归分析，求出此时因变量与新的自变量之间的 函数模型即可

八、模型的改进与推广

模型一建立的假设是所有因变量取值必须是连续的，而公司运营状态是变化的，这就无法进行保证，下面在模型一的基础上对其进行改进推广，建立）偏最小二乘$\log istic$回归模型

8.1 偏最小二乘$\log istic$回归模型

算法步骤：

记 $Y=[y_1,y_2,\ldots y_n]^T$， $y_{i(i=1,2,\ldots n)}$ 在值域 $\{1,2,\ldots c\}$ 取值

$X=\left(x_{ij}\right)_{n\times p}$ 为标准化自变量矩阵

其中有 p 个解释变量 $x_{j(j=1,2,\ldots p)}$

一、 提取偏最小二乘成分

（1） 提取第一个偏最小二乘成分 t_1

第一步：

分别建立因变量 Y 对自变量 $x_{j(j=1,2,\ldots p)}$ 的普通 $\log istic$ 回归模型，在模型中， $x_{j(j=1,2,\ldots p)}$ 的回归系数为 $\omega^*_{1j(j=1,2,\ldots p)}$

第二步：

$$\omega^*_1=\left(\omega^*_{11},\omega^*_{12},\ldots\omega^*_{1p}\right)^T$$

将 ω^*_1 标准化得 ω_1

$$\omega_1=\left(\omega_{11},\omega_{12},\ldots\omega_{1p}\right)^T$$

$$\omega_{1j}=\omega^*_{1j}\Big/\sqrt{\sum_{j=1}^{p}\omega_{1j}^{*2}}\left(j=1,2,\ldots p\right)$$

第三步：

提取第一个成分 t_1，有

$$t_1=\frac{x\omega_1}{\left\|\omega_1\right\|^2}=\frac{x\omega_1}{\omega_1^T\omega_1}$$

（2） 提取第二个偏最小二乘成分 t_2

第一步：

建立 X 对第一个成分 t_1 的回归模型

$$X=t_1p_1^T+X_1$$

其中， p_1 是回归系数，即 $p_1=\frac{X^Tt_1}{\left\|t_1\right\|^2}$，得到 X 的残差矩阵 X_1 且记 $X_{1j(j=1,2,\ldots p)}$ 为残差矩阵 X_1 的第 j 列

第二步：

对每一个 $j\left(j=1,2,\ldots p\right)$ 分别建立因变量 Y 对 t_1， X_{1j} 的 $\log istic$ 回归

模型，并记ω_{2j}^{*}为模型X_{1j}的回归系数

第三步：

$\omega_2^{*}=\left(\omega_{21}^{*},\omega_{22}^{*},\ldots\omega_{2p}^{*}\right)^{T}$

$\omega_2=\left(\omega_{21},\omega_{22},\ldots\omega_{2p}\right)$

第四步：

计算第二个成分t_2

$$t_2=\frac{X_1\omega_2}{\left\|\omega_2\right\|^2}=\frac{X_1\omega_2}{\omega_2^{T}\omega_2}$$

（3） 在已提取$(h-1)$个偏最小二乘成分$t_1,t_2,\ldots t_{h-1}$后提取第h个偏最小二乘成分t_h

第一步：

建立X对$t_1,t_2,\ldots t_{h-1}$的回归模型

$$X=t_1p_1^{T}+t_2p_2^{T}+\ldots+t_{h-1}p_{h-1}^{T}+X_{h-1}$$

其中$p_{k(k=1,2,\ldots h-1)}$是回归系数，即$p_k=\frac{X_{k-1}^{T}t_k}{\left\|t_k\right\|^2}(k=1,2,\ldots h-1)$得残差矩阵$X_{h-1}$

并记$X_{h-1,j(j=1,2,\ldots p)}$为矩阵$X_{h-1}$的第$j$列向量

第二步：

对每个$j(j=1,2,\ldots p)$，分别建立因变量Y对$t_1,t_2,\ldots t_{h-1}$和$X_{h-1,j}$的$\log istic$回归模型，并记ω_{hj}^{*}为模型中X_{h-j}的回归系数。

第三步：

将列向量$\omega_h^{*}=\left(\omega_{h1}^{*},\omega_{h2}^{*},\ldots\omega_{hp}^{*}\right)^{T}$进行标准化，并记为$\omega_h$

第四步：

提取成分t_h，有$t_h=\frac{X_{h-1}\omega_h}{\left\|\omega_h\right\|^2}=\frac{X_{h-1}\omega_h}{\omega_h^{T}\omega_h}$

第五步：

将t_h表示为原始变量的线性组合形式，有$t_h=X\varpi_h$

其中 $\varpi_h = \prod_{k=1}^{h-1}\left(E - \omega_k p_k^{\ t}\right)\omega_h$

二、 建立 $\log istic$ 模型

建立因变量 Y 对成分 $t_1, t_2, \ldots t_n$ 的 $\log istic$ 回归模型

$$\ln\left(\frac{p\left(y \le (t_1, t_2, \ldots t_n)\right)}{1 - p\left(y \le (t_1, t_2, \ldots t_n)\right)}\right) = \beta_{oc} + \sum_{k=1}^{h}\beta_k t_k = \beta_{oc} + \sum_{k=1}^{h}\beta_k X \varpi_k = \beta_{oc} + \sum_{j=1}^{p} b_j X_j$$

其中 $b_j = \sum_{k=1}^{h}\beta_j \varpi_{kj}$

三、 提取成分的标准

判断选取偏最小二乘成分个数 h 的标准，主要采用各回归系数的统计显著性检验，拟合优度指标 AIC ， SC ， $-2LL$ 值以及预测错误等，使得当 $h = ?$ 时，回归方程显著。

在原有数据的基础上将数据代入 $\log istic$ 回归模型即可求出相应的指标饱和估计值，得出的数值更加准确可靠。

九、季度报告

（1）根据模型一多元回归模型和改进后的 $\log istic$ 回归模型发现业务 2 和业务 3 已经接近饱和状态，而业务 1、业务 4、业务 5 仍有发展趋势，尤其是业务 4 发展势头强劲。作为经理，为了公司的长远发展，接下来应该大力发展业务 4，把其打造成公司的特色品牌，重点推进业务 1、业务 5，但是这并不意味着我们要放弃业务 2 和业务 3，相反需要采取措施赢得顾客需求，比如明星效应、广告宣传等等

（2）根据模型二相关系数模型得出的相关系数值大小知道，收入主要和业务 2、业务 3 和业务 4 相关而与业务 5 无关。公司是一个整体，公司追求的是整体的经济效益，公司需要个别优秀业务的发展带动其他业务的发展，这样才有利于公司的长久生存，通过模型三我们能够确定哪些业务之间有相互促进的关系哪些则没有，继而在接下来的一段时间内我们着重发展这些业务，同时创造环境使得各个业务都能够参与其中共同个发展。

十、参考文献

[1]胡运权主编 郭耀辉副主编，运筹学教程（第二版 ），北京:清华大学出版社和，2003
[2]陈珽，决策分析，北京:科学出版社，1987
[3]堵秀凤、张剑、张宏民编著，数学建模，北京:北京航空航天大学出版社，2011
[4]姜启源，数学模型（第二版），北京:高等教育出版社,2003
[5] 李志林，欧宜贵编，《数学建模及典型案例分析》，北京:化学工业出版社，2006.12

[6]王惠文，吴载斌，孟洁著，偏最小二乘回归的线性与非线性方法，北京：国防工业出版社，2006.9.
[7]陆元鸿编著，数理统计方法，上海：华东理工大学出版社，2005.8.
[8]茆诗松等编著,回归分析及其试验设计,上海:华东师范大学出版社,1981.10.
[9]曾五一 肖红叶主编，统计学导论，北京：科学出版社，2007.1

附录

附录一

日期	日期	指标 1（单位万）
2012/1/1	1	12,500
2012/1/2	2	13,503
2012/1/3	3	14,872
2012/1/4	4	15,092
2012/1/5	5	15,001
2012/1/6	6	14,098
2012/1/7	7	14,001
2012/1/8	8	13,520
2012/1/9	9	14,097
2012/1/10	10	15,011
2012/1/11	11	15,020
2012/1/12	12	15,101
2012/1/13	13	14,073
2012/1/14	14	13,501
2012/1/15	15	12,976
2012/1/16	16	13,509
2012/1/17	17	13,076
2012/1/18	18	12,187

2012/1/19	19	10,655
2012/1/20	20	9,276
2012/1/21	21	8,012
2012/1/22	22	7,032
2012/1/23	23	8,132
2012/1/24	24	8,298
2012/1/25	25	9,076
2012/1/26	26	9,875
2012/1/27	27	10,598
2012/1/28	28	11,045
2012/1/29	29	12,098
2012/1/30	30	12,765
2012/1/31	31	14,076
2012/2/1	32	15,576
2012/2/2	33	16,843
2012/2/3	34	16,432
2012/2/4	35	15,987
2012/2/5	36	15,102
2012/2/6	37	16,908
2012/2/7	38	17,098
2012/2/8	39	17,398
2012/2/9	40	17,987

2012/2/10	41	18,654
2012/2/11	42	16,098
2012/2/12	43	15,806
2012/2/13	44	17,320
2012/2/14	45	16,987
2012/2/15	46	17,298
2012/2/16	47	18,332
2012/2/17	48	19,087
2012/2/18	49	17,098
2012/2/19	50	16,543
2012/2/20	51	18,654
2012/2/21	52	18,987
2012/2/22	53	20,543
2012/2/23	54	20,076
2012/2/24	55	21,067
2012/2/25	56	19,168
2012/2/26	57	18,643
2012/2/27	58	20,321
2012/2/28	59	21,009
2012/2/29	60	20,877
2012/3/1	61	20,575
2012/3/2	62	19,865

2012/3/3	63	19,321
2012/3/4	64	18,765
2012/3/5	65	21,586
2012/3/6	66	21,087
2012/3/7	67	20,986
2012/3/8	68	20,487
2012/3/9	69	20,746
2012/3/10	70	18,735
2012/3/11	71	18,328
2012/3/12	72	20,433
2012/3/13	73	20,986
2012/3/14	74	21,076
2012/3/15	75	22,376
2012/3/16	76	22,508
2012/3/17	77	20,875
2012/3/18	78	19,165
2012/3/19	79	22,583
2012/3/20	80	22,683
2012/3/21	81	21,876
2012/3/22	82	23,543
2012/3/23	83	21,976
2012/3/24	84	18,654

2012/3/25	85	18,708
2012/3/26	86	22,098
2012/3/27	87	22,598
2012/3/28	88	22,865
2012/3/29	89	23,376
2012/3/30	90	22,875
2012/3/31	91	21,987