# 艾滋病药物治疗的研究

## 摘要

本文研究了艾滋病疗效的预测以及疗法的评价问题。首先我们对数据进行了分析和处理，排除了相应的缺失数据和不匹配数据，通过Spearman秩相关分析，研究了数据之间的相关关系。

在基本问题的求解中，我们做出合理假设，简化关系。针对问题一，我们定义了衡量模型优劣程度的平均相对误差，讨论了终止治疗的最佳时间，分别基于正态分布，多项式回归预测，支持向量机预测建立了三个模型。使用开放源代码的 Matlab 工具包 LibSVM 来执行支持向量机算法,得到的表达式从拟合度和稳定性两方面考虑都较为令人满意，由该算法我们进行了问题一的求解。

对于问题二，我们将CD4数目的增加值作为评价疗法优劣的指标，通过统计相关数据，定义了评价 4 种疗法优劣程度的效用函数g，得出疗法 4 是相对较优的疗法，并利用支持向量机算法确定了最佳终止治疗时间。

在问题三中，我们在第二问的基础上引入了价格因素$p^r$，由费用敏感指数r来区别不同人对治疗费用敏感程度的差异，定义了性价比函数U，综合考虑费用和疗效对疗法的影响，并求解出费用敏感指数r=1时疗法 1 相对较优。同时，利用支持向量机的算法确定了疗法 1 的最佳终止时间。

在模型拓展中，我们放松假设，逼近现实，对停药时间和预测方法进行了进一步讨论。

关键词：Spearman秩相关系数，平均相对误差 s，支持向量机算法，效用函数g，费用敏感指数 r，性价比函数U

# 目　录

# 1．问题分析

1．艾滋病的发病机理：

艾滋病的医学全称为“获得性免疫缺损综合症”，它是由艾滋病病毒（HIV）引起的。在人体的免疫系统中，CD4细胞在抵御HIV入侵的最主要的免疫细胞。而HIV病毒体主要袭击的目标是CD4细胞，它通过渗入CD4细胞膜，控制CD4的繁殖器官，把CD4作为HIV病毒的“加工厂”，进行大量的繁殖。所以，当人体受到HIV病毒攻击时，CD4细胞会急剧减少，HIV病毒的数量会迅速增加，导致了艾滋病发作。

2．艾滋病的治疗：

一旦被确诊为艾滋病患者，就要接受一些治疗。目前主要是药物治疗，但当患病者长期用一种药时，HIV就会发生变异，从而对该药物产生了抗药性，因此会出现病情反复的情况。所以采用联合疗法（即使用多种药物治疗艾滋病）比单一疗法（只用一种药治疗艾滋病）的效果要好。根据附件1所给数据，研究CD4和HIV的相关性，当继续服药的效果不好时，则可选择提前终止治疗。那么就需要确定什么时候为最佳的终止治疗时间。

3．评价疗法的效果：

目前治疗艾滋病有多种疗法，但治疗的效果却有一些差异，有些疗法还会产生副作用。治疗爱滋病的目的就是尽量减少HIV的数目，产生更多的CD4。那么我们可以以CD4为标准，根据附件2所给数据，找出评价疗法优劣的指标函数，对题目中的四种疗法进行评价。另外，在治疗的过程中不仅要考虑到疗效，还要考虑治疗费用，就需要综合两方面因素，来制定评价疗法优劣的指标。

# 2．问题重述

艾滋病自1981年在美国发现以来，就像一个恐怖的“黑色幽灵”，在全球范围内无节制地广为流行。至今，它已夺去近300万人的生命。

艾滋病的医学全名为“获得性免疫缺损综合症”，英文简称AIDS，它是由艾滋病毒（医学全名为“人体免疫缺损病毒”，英文简称HIV）引起的。人类免疫系统的CD4细胞在抵御HIV的入侵中起着重要作用，而HIV病毒体主要袭击的目标是CD4细胞，它通过渗入CD4细胞膜，控制CD4的繁殖器官，把CD4作为HIV病毒的“加工厂”，进行大量的繁殖，导致AIDS发作。

现在治疗艾滋病的疗法很多，有抗感染治疗、抗病毒治疗、抗肿瘤治疗、免疫调节及免疫重建治疗等，但迄今为止人类还没有找到能根治AIDS的疗法，目前的一些AIDS疗法不仅对人体有副作用，而且成本也很高。

现在得到了美国艾滋病医疗试验机构ACTG公布的两组数据。 ACTG320（见附件1）是同时服用zidovudine（齐多夫定）， lamivudine（拉美夫定）和indinavir（茚地那韦)3种药物的300多名病人每隔几周测试的CD4和HIV的浓度(每毫升血液里的数量)。193A（见附件2）是将1300多名病人随机地分为4组，每组按下述4种疗法中的一种服药，大约每隔8周测试的CD4浓度（这组数据缺HIV浓度，它的测试成本很高）。4种疗法的日用药分别为： 600mg zidovudine或400mg didanosine（去羟基苷），这两种药按月轮换使用； 600 mg zidovudine加2.25 mg zalcitabine （扎西他滨）；
600 mg zidovudine加400 mg didanosine； 600 mg zidovudine加400 mg didanosine，再加400 mg nevirapine（奈韦拉平）。

我们将完成以下问题：

（1）利用附件1的数据，预测继续治疗的效果，或者确定最佳治疗终止时间（继续治疗指在测试终止后继续服药，如果认为继续服药效果不好，则可选择提前终止治疗)。

（2）利用附件2的数据，评价4种疗法的优劣（仅以CD4为标准)，并对较优的疗法预测继续治疗的效果，或者确定最佳治疗终止时间。

(3) 艾滋病药品的主要供给商对不发达国家提供的药品价格如下：600mg zidovudine 1.60美元，400mg didanosine 0.85美元，2.25 mg zalcitabine 1.85美元，400 mg nevirapine 1.20美元。如果病人需要考虑4种疗法的费用，对（2）中的评价和预测（或者提前终止）有什么改变。

## 3．模型假设

1. 假设在人体的免疫系统中，只有CD4能抵御HIV病毒的入侵，不考虑其它免疫细胞的免疫作用；
2. 假设CD4的数目和HIV的浓度能完全反映治疗的效果，忽略其它因素的影响；
3. 忽略病情出现反复的情况；
4. 假设病人均为第一次接受某种疗法，不存在该次治疗前产生的抗药性的影响；
5. 由于抗药性的影响，每种疗法治疗一段时间后，疗效会显著下降，治疗终止仅意味着一种治疗方法的终止，病人可改用其它的疗法。

## 4．符号说明

$C^1$： 所有接受测试的病人的CD4的数目所组成的列向量(个/$mm^3$)；

$H^1$： 所有接受测试的病人的 HIV的浓度所组成的列向量(有部分数据缺失)；

$C^1_{i,k}$： 第i个病人第k次测试时CD4的数目(个/$mm^3$)；

$H_{i,k}^1$：　第i个病人第k次测试时 HIV的浓度(单位不详)；

$C_i(t)$：　$t$时刻所有病人i的CD4数目所组成的列向量(个/$mm^3$)；

$t_i$：　表示第i个病人的最佳治疗终止时间(周)；

$\rho_1$：　表示CD4数目和 HIV浓度的Spearman秩相关系数；

g：　效用函数，用来评价只考虑疗效时的疗法的优劣；

U：　性价比函数，即考虑疗效和费用的疗法综合评价指标, U值越大, 疗法越优；

r：　病人对于费用的敏感指数， r越大，病人对费用越敏感；

p：　某种疗法的费用(美元)；

q：　疗法的失效率，即某种疗法治疗无效(CD4数目没有增加)的病人数占该疗法总人数的百分比；

如用到其它符号，在文中予以说明。

# 5．模型建立与求解

## 5.1 问题一：

### 5.1.1 数据观察与预处理：

初步观察数据，可以发现随着治疗的持续，CD4数目先是增加，然后减少；HIV浓度则正好相反。这说明在治疗一段时间后，在CD4数目不见增加或者HIV浓度不见减少时应终止治疗。从给出的大量样本数据，通过统计方法可以预测如果继续治疗，CD4数目和HIV浓度将怎样变化。

在附件 1 的数据中，有一部分缺失值。CD4数目的数据有 3 处缺失；HIV浓度的数据有 93 处缺失（这可能是因为此项测试成本较高的原因）。在进行基于CD4数目或HIV浓度的分析时，我们将排除相应的缺失数据。同时我们还注意到，即使一条记录中CD4数目和HIV浓度的数据都存在，它们的测试时间并不一定相等，有 14 条这样的测试时间不匹配的数据。我们把不含缺失值且测试时间匹配的数据统称为匹配数据。对于个别病人不从第 0 周开始的测试，我们将日期平移调整到第 0 周开始。

由背景知识可知，CD4的数目向量（记作$C^1$）和HIV的浓度向量（记作$H^1$）有一定的关系，但并不见得是线性关系。为了考察这种关系，我们首先对$C^1$和$H^1$作非参数相关分析[1]，计算它们的Spearman秩相关系数$\rho_1$。与线性相关系数比较，秩相关系数仅考察了两变量大小变化是否有关，而不考察变量之间的线性关系。相关分析的对象为为匹配数据，计算得到$\rho_1 = -0.4317$。

与Spearman秩相关系数统计量的临界值比较，$|\rho_1|$值偏大，因此$C^1$和$H^1$并不独立。

而 $\rho_1 < 0$ 印证了“当CD4被HIV感染而裂解时，其数量会急剧减少，HIV将迅速增加”的观点。但是 $|\rho_1|$ 仍离 1 较远，说明 $C^1$ 和 $H^1$ 的关系并不明显。两者的散点图也说明了这一点（见图 1）。考虑到病人开始治疗时（第 0 周，第 1 次测试）的身体情况以及病人个体条件会不一致，出现这样的结果应该是预料之中的。为了消除病人初始状态的影响，我们对 $C^1$ 和 $H^1$ 作差分处理。

设接受治疗和检查的人数为 $N_1$，我们对 $C^1$ 作差分

$$C^2_{i,\mathrm{k}} = C^1_{i,\mathrm{k}} - C^1_{i,\mathrm{k}-1} \ (\mathrm{k} > 1, i = 1,..., N_1)\ ,$$

将 $C^2_{i,\mathrm{k}}$ 按顺序排成列向量 $C^2$。与 $C^1$ 相比，$C^2$ 的长度比 $C^1$ 减少了 $N_1$。对于 $H^1$ 我们也作类似处理得到 $H^2$。此外，对于病人 $i$，第 1 次测试和第 2 次测试之间的间隔时间计为 $d_i$，

$$d := (d_1, d_2, ...d_{N_1})$$

若CD4和HIV互斥的规律严格成立，则 $C^2$ 和 $H^2$ 的符号应该总是相反的。而对比可见样本中它们约有 25%~38%的符号不一致（部分样本中它们的数值为 0，故无法确定），这说明了CD4数目和HIV浓度并没有必然的关系，有必要对它们分别进行分析。

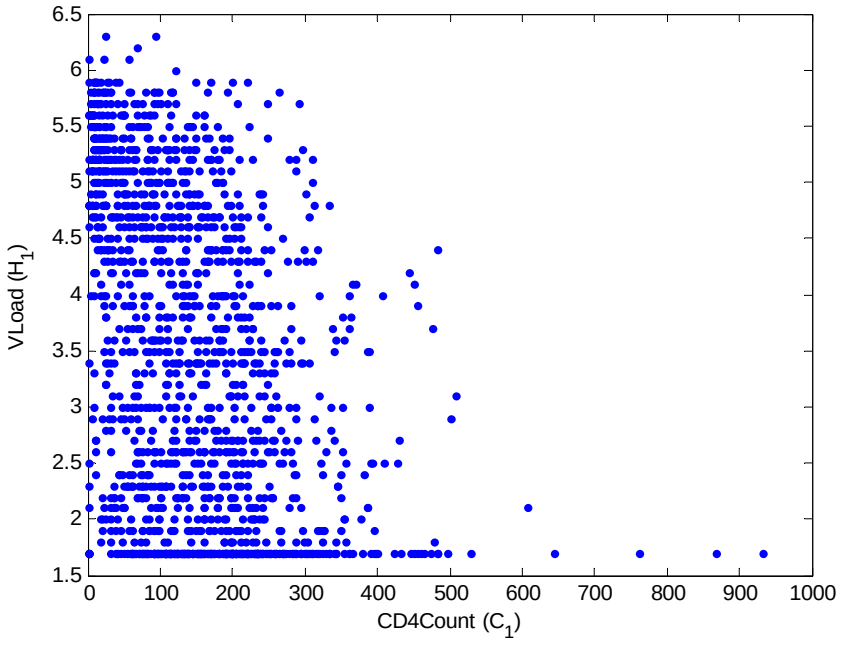


图 1

5.1.2 确定最佳治疗终止时间的预测模型：

5.1.2.1 模型准备：

首先我们来确定终止治疗的时间，由于抗药性的存在，每种疗法治疗一段时间后，疗效会显著下降，把CD4数目作为终止治疗的依据，即当CD4恰达到最大值时，终止治疗。我们的算法应当预测一个病人该何时终止治疗，以免继续治疗下去没有好的疗效。假设病人 $i(i = 1,..., N_1)$ 应当终止治疗的时间为 $t_i$，算法得到的终止治疗时间为 $\hat{t}_i$，衡量算法好坏的标准是如下定义的平均相对误差：

$$s=\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{|t_i-\hat{t}_0|}{\sqrt{\mathrm{var}(t)}}$$

其中 $\sqrt{\mathrm{var}(t)}$ 表示样本 $t_i(i=1,...,N_1)$ 的标准差，使用它而不用通常的 $t_i$ 是因为可能出现 $t_i=0$ 的情形。

注意到有所谓“病情反复”的情况，即在治疗过程中，CD4 先是增加，然后不再增加（此时本应终止治疗），然后又重新增加达到某个较大的值。这时终止治疗时刻的定义是不明确的。

在下面的两个模型中，遇到“病情反复”时，我们考虑的是取最大 CD4 值的时刻，因为这是治疗所能达到的最好程度。有些病人的 CD4 在治疗期内一直增加，我们把他们的终止治疗时间定为最后一个测试时刻。

5.1.2.2 基于正态分布的预测终止时间的模型：

假设终止治疗的时间与病人无关。即每个病人在经过时间 $t_0+\varepsilon$ 的治疗后，就不会再增加 CD4。其中 $t_0$ 是一个常数，$\varepsilon$ 是随机误差。

按照前述确定终止治疗时间的原则，我们从样本数据获得了每个病人终止治疗的时间：

$$t_i(i=1,...,N_1)=t_0+\varepsilon_i$$

除去缺失值。假设 $\varepsilon_i$ 服从正态分布，得到 $t_0$ 的估计 $\hat{t}_0=29.0$。即对所有的病人，在治疗 29 周后终止治疗。

通过计算可以得出，样本 $t_i(i=1,...,N_1)$ 的标准差为 13.1，平均相对误差 $s$ 高达 86.75%。这样看来模型 1 得到的结果是没有多少意义的。我们必须改进这个模型，考虑不同的病人条件对终止治疗时间的影响。

5.1.2.3 基于 $k-$最近邻法和支持向量机预测终止时间的模型：

假设终止治疗的时间与病人初始状态和病人接受治疗的能力有关。病人 $i$ 的初始状态即 $C_{i,1}^1$ 和 $H_{i,1}^1$；他接受治疗的能力用开始治疗阶段每天 CD4 的增加数来表示，即 $C_{i,1}^2/d_i$ 和 $H_{i,1}^2/d_i$。我们的模型可以描述为：

$$t_i=f(\mathrm{C}_{\mathrm{i},1}^1,H_{i,1}^1,C_{i,1}^2/d_i,H_{i,1}^2/d_i)+\varepsilon_i$$

其中 $\varepsilon_i$ 是针对病人 $i$ 的随机误差，$f(\cdot)$ 是一个函数，我们假设它是连续函数。

在没有关于 $f(\cdot)$ 的更多信息的情形下，我们不能对它假定像线性模型这样的结构，一般也不能显式的写出 $f(\cdot)$ 的估计。但是我们仍能通过一些算法，从样本中获得关于 $f(\cdot)$ 的信息，从而对于病人 $j$，从他的 $C_{j,1}^1$，$H_{j,1}^1$，$C_{j,1}^2/d_i$ 和 $H_{j,1}^2/d_i$ 来预测 $t_j$，从而达到确定终止治疗的时间的目的。在这里，我们讨论 $k-$最近邻法和支持向量机这两种算法，它们都能满足我们的要求。

$k-$最近邻法是一种经典的非参数方法，关于它的详细说明可以从文献[2]中找到，这里限于篇幅不再赘述。

支持向量机是新近出现的回归算法，从同名的分类算法修改而来[2]。我们使用开放源代码的 Matlab 工具包 LibSVM 来执行支持向量机[4]。得到的两种算法的评价误差如

表一所示：

| 方法 | 线性回归 | $k-$最近邻法 | 支持向量机 |
|---|---|---|---|
| 平均相对误差 | 97.68 % | 78.35 % | 18.38 % |

表 1

从表 1 可以看出，$k-$最近邻法的效果不佳，说明因为函数 $f(\cdot)$ 的变化过快，邻近点的函数值的平均不能很好的逼近真实值。而支持向量机明显好于其它算法，这是因为它能够逼近任意连续函数的缘故。作为对比，我们还给出了线性回归的结果。

尽管支持向量机对已有样本的效果不错，但由于 $f(\cdot)$ 本质上的不稳定，导致该算法不稳定。如果抛弃某一个样本点后重新计算支持向量机，并应用到被抛弃样本上，误差仍比较大。下面我们换一个角度来确定终止治疗的时间。

5.1.2.4 基于多项式回归预测CD4数目的模型：

我们知道，由于抗药性的存在，药物在经过一段时间后会渐渐失效，CD4数目的增加会越来越慢，直到开始变少。把CD4数目 $C^1$ 看作时间 $t$ 的函数 $C(t)$，则 $dC(t)/dt$ 在 $t$ 较小时为正，随着 $t$ 变大时越来越小，直到变成负数。但是由于各种药物抗药性出现的时间不一致，还会发生“病情反复”的情况。此时最佳的治疗终止时间为 CD4 含量开始下降的时间。

在文献[5][6]中，对于 $C(t)$ 应当服从的微分方程作了讨论。它的形式比较复杂，要定量的确定参数值不是一件容易的事。在这里，我们先忽略病情反复的情况，假定

$$\frac{dC(t)}{dt} = at + b \quad (a < 0)$$

解得

$$C(t) = \frac{a}{2}t^2 + bt + c$$

对于不同的病人，参数 $a,b,c$ 是不同的，它们应该跟病人的初始状态和接受治疗的能力有关，即

$$a_i = a(\mathrm{C}^1_{i,1}, H^1_{i,1}, C^2_{i,1}/d_i, H^2_{i,1}/d_i)$$

$$b_i = b(\mathrm{C}^1_{i,1}, H^1_{i,1}, C^2_{i,1}/d_i, H^2_{i,1}/d_i)$$

由于 $C(0) = C^1_{i,1}$，考虑到随机误差，取 $c_i = C^1_{i,1} + \varepsilon$，$\varepsilon$ 是随机误差。

考虑函数 $a(\cdot), b(\cdot)$ 均为线性的情形，问题转化为一个多项式回归的问题。求解得到回归方程为

$$C_i(t) = (2.56 - 0.00930\mathrm{C}^1_{i,1} - 0.0.453H^1_{i,1} - 0.0128C^2_{i,1}/d_i + 0.194H^2_{i,1}/d_i)t^2$$

$$+ (-36.0 + 0.234\mathrm{C}^1_{i,1} + 6.395H^1_{i,1} + 0.475C^2_{i,1}/d_i - 8.01H^2_{i,1}/d_i)t + C^1_{i,1}$$

对于病人 $i$，解方程 $C_i'(t)=0$，可以获得最佳的终止治疗时间 $t_i$。然而用这个方程得

到的平均相对偏差为 65.7%，仍然不够理想。这说明 $a(\cdot),b(\cdot)$ 不是简单的线性关系，我们有必要考虑非参数的方法。

5.1.2.5 基于支持向量机预测 CD4 数目的方法：

我们这里只是假设连续函数，且对每个病人 i：

$$C_{\mathrm{i}}(t)=C(\mathrm{C}_{\mathrm{i},1}^{1},H_{i,1}^{1},C_{i,1}^{2}/d_i,H_{i,1}^{2}/d_i,\mathrm{t})$$

我们仍利用支持向量机来求出函数 $C_{\mathrm{i}}(t)$。为了验证稳定性，我们把样本分为两部分，一部分用来训练得到 $C_{\mathrm{i}}(t)$，一部分用来测试衡量误差。我们得到在测试集上的平均相对偏差为 30.7%，支持向量机拟合得到的函数关系式从拟合度和稳定性两方面考虑都较为令人满意。

但是由于 $C_{\mathrm{i}}(t)$ 没有显示的表达式，终止时间的确定，即方程 $C_i'(t)=0$ 无法解析求解。但我们可以通过数值微分的方法得到。且在存在多个解使得 $C_i'(t)=0$ 时取第一个时间。求得的结果如表 2 所示（仅随机列出 10 个病人）。

考虑到测试做到了第 40 周，每隔 8 周一次，故预测第 48 周的疗效(下同)。

| PtID | 终止时间 | 预测疗效（第 48 周）<br>CD4 |
|---|---|---|
| 23441 | 9 | 152 |
| 23442 | 15 | 217 |
| 23443 | 16 | 157 |
| 23444 | 20 | 145 |
| 23445 | 13 | 139 |
| 23447 | 39 | 176 |
| 23448 | 24 | 120 |
| 23449 | 19 | 102 |
| 23450 | 26 | 72 |
| 23451 | 32 | 84 |

表 2

注：ID23446 中，HIV 浓度的数据缺失，基于我们前述的处理方法，已排除了相应的缺失数据。

上述问题我们仅考虑了 CD4 数值这一个指标。当然我们也可以考虑 HIV 浓度，或者把两者综合考虑。只需要把上述问题中的目标函数由 CD4 的数目修改为 HIV 的浓度，或者 CD4 数目/HIV 浓度即可，这样同样可以利用模型 3 的支持向量机算法。

我们以 CD4 数目/HIV 浓度作为目标，可以得到表 3 的结果(不妨取表 2 中的 10 个病人)。

| PtID | 终止时间 | 预测疗效（第 48 周）CD4/HIV |
|---|---|---|
| 23441 | 10 | 24.3 |
| 23442 | 13 | 38.9 |
| 23443 | 12 | 48.5 |
| 23444 | 20 | 63.8 |
| 23445 | 22 | 64.2 |
| 23447 | 35 | 52.3 |
| 23448 | 21 | 32.8 |
| 23449 | 46 | 95.5 |
| 23450 | 26 | 57.6 |
| 23451 | 27 | 20.9 |

表 3

注：ID23446 中；HIV 浓度的数据缺失，基于我们前述的处理方法，已排除了相应的缺失数据。

## 5.2 问题二：

5.2.1 数据处理：

通过观察，我们注意到附件 2 数据中没有直接给出CD4数目，我们需要作一变换：我们认为附件 2 的中所给的CD4数目是实际CD4的值减 1 后取了以 e 为底的对数。我们按照$e^x-1$（x为原始数据）公式进行还原得到实际CD4的值，记为C（所有病人所组成的列向量）。对于只参加了一次测试的病人，无法分析药的疗效，我们将这些数据删去，得到剩余病人的CD4的数目向量$C^3$。

经过Spearman秩相关检验，我们发现CD4数目的增加和病人的年龄以及CD4数目的初始值没有关系。因此这些因素在比较用药方案的优劣时不作考虑。

四种疗法所对应的药物说明见表 4。

| 疗法 | 所对应的药物 |
|---|---|
| 1 | 600mg zidovudine 或 400mg didanosine （去羟基苷），按月轮换使用 |
| 2 | 600 mg zidovudine , 2.25 mg zalcitabine （扎西他滨） |
| 3 | 600 mg zidovudine , 400 mg didanosine |
| 4 | 600 mg zidovudine , 400 mg didanosine , 400 mg nevirapine （奈韦拉平） |

表 4

5.2.2 艾滋病疗法的优劣评价模型：

在这里我们把CD4数目作为评价用药方案好坏的指标。对于某种疗法的某个的病人来说，他的CD4增加得越多，说明该疗法越有效。我们定义病人$i$接受某种疗法后的CD4数目增加值为：

$$D_i = C_{i,m}^3 - C_{i,1}^3$$

其中$C_{i,m}^3$是接受该种疗法的病人$i$在几次测试中CD4达到的最大值，如果$D_i > 0$说明使用该种疗法是有效果的。设接受该疗法治疗的总人数为N，而有治疗效果的人数为n；而$D_i$=0说明该疗法对于病人i是没有效果甚至是有副作用的，因此对于$D_i = 0$另作统计（即疗法的失效率，记为q）。我们对得到的$D_i > 0$的数据进行统计，利用以下公式：

$$\text{平均值：}\ \overline{\mathrm{D}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{D_i}$$

$$\text{标准差：}\ \delta = \sqrt{\frac{1}{\mathrm{n\text{-}1}}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{D_i} - \overline{\mathrm{D}})^2}$$

$$\text{失效率：}\ \mathrm{q} = \frac{\mathrm{N} - \mathrm{n}}{\mathrm{N}}$$

分别计算出四种疗法对于有疗效的病人（$D_i > 0$）CD4增加数目的平均值、标准差和失效率（见表3）。$\overline{\mathrm{D}}$反映了疗法j的平均治疗效果，$\overline{\mathrm{D}}$越大，即CD4细胞增加的越多，说明疗效越好；$\delta$反映了疗法的稳定程度，$\delta$越小，说明$D_i$的波动越小，越稳定；p越小，说明该疗效产生的副作用越小，说明疗效越好。根据以上分析，我们定义疗法的效用函数：

$$\mathrm{g} = \frac{\overline{\mathrm{D}} \cdot (1 - \mathrm{p})^2}{\delta}$$

由定义式可以看出，g越大，疗法的疗效就越好。

| 疗法 | 疗法1 | 疗法2 | 疗法3 | 疗法4 |
|---|---|---|---|---|
| $\overline{D}$ | 19.2741 | 24.4595 | 38.5680 | 45.0597 |
| $\delta$ | 28.5966 | 29.3831 | 53.5344 | 69.0220 |
| q | 49.5 | 39.8% | 41.5% | 23.7% |
| g | 0.171887 | 0.301678 | 0.246551 | 0.380058 |

表 5

由表5可以看出，根据我们定义的效用函数，对4种疗法进行优劣次序为：疗法4>疗法2>疗法3>疗法1。显而易见，疗法4是最优的。

CD4数目的增加量还可以有别的定义。一种是利用问题1中的方法，确定出最优终止时间后，最优时间达到的最大值与初始值的差；另一种是利用问题1中的方法，预测经过一段时间的治疗后（如48周）CD4能达到的值与初始值的差。这两种定义得到的结果与前述定义得到的结果类似，不再赘述。

我们用问题一中效果最好的模型，即模型3的支持向量机算法，来对采取疗法4的病人确定终止治疗时间和预测疗效，其结果如表6(仅随机选取10个病人)。

| PtID | 终止时间 | 预测疗效（第 48 周）<br>按前述方法处理后的值 |
|---|---|---|
| 2 | 19 | 14.7 |
| 12 | 14 | 18.4 |
| 26 | 20 | 20.1 |
| 29 | 18 | 21.8 |
| 31 | 15 | 6.8 |
| 36 | 10 | 4.2 |
| 46 | 13 | 24.5 |
| 50 | 28 | 21.4 |
| 55 | 15 | 17.3 |
| 62 | 12 | 14.9 |

表 6

## 5.3 问题三：

### 5.3.1 模型准备：

我们根据题目中给出了四种药品的价格(表 7 ),计算出四种疗法的治疗费用，其中疗法 1 由于是两种药品按月轮流使用，我们取两种药品价格的平均值来作为疗法 1 的治疗费用，其它疗法费用是各药品价格的和，如表 8 所示。

| 药物 | 药物的费用（美元 ） |
|---|---|
| 600mg zidovudine | 1.60 |
| 400 mg didanosine | 0.85 |
| 2.25 mg zalcitabine | 1.85 |
| 400 mg nevirapine | 1.20 |

表 7

| 疗法 | 所对应的费用（美元） |
|---|---|
| 1 | 1.225 |
| 2 | 3.45 |
| 3 | 2.45 |
| 4 | 3.65 |

表 8

### 5.3.2 模型的建立与求解：

如果病人考虑到四种疗法的费用，结合自身的经济承受能力来选择疗法。我们引入了价格因素 $p^r$，其中 p 是疗法所需的费用，r（r>0）是病人对费用的敏感指数。由于个人

经济承受能力的不同，每个人对费用的敏感程度是不同的，因此，我们在模型中引入了费用敏感指数，r 越大，说明病人对费用越敏感。当 r=0（即 $p^r=1$）时，价格因素不起作用，病人对费用不敏感，表示病人只追求最好的治疗效果，无论付出多大的费用。当 r=1（即 $p^r=p$）时，价格因素对病人选择的疗法有较大的影响，此时病人会折中考虑费用和疗效。可能有的疗法效果稍差一些，但费用较低，有些对费用比较敏感的病人愿意出较少的费用来获得稍差的治疗效果。而在相同治疗效果的情况下，费用越低，则该疗法越好。综合考虑效果和费用的共同作用，我们建立了如下的性价比函数 U 来评价疗法的优劣：

$$U=\frac{g}{p^r}$$

其中 g 与问题二中定义相同。根据综合性价比函数的定义式，U 越大，疗效越好。

模型的求解结果见表 9

| 用药方案 | 疗法 1 | 疗法 2 | 疗法 3 | 疗法 4 |
|---|---|---|---|---|
| U(r=0) | 0.171887 | 0.301678 | 0.246551 | 0.380058 |
| U(r=1) | 0.140316 | 0.0874428 | 0.100633 | 0.104125 |

表 9

若病人对费用不敏感，即当 r=0 时，第 4 种疗法的性价比函数 U 最大，所以对于疗法费用不敏感的人来说，疗法 4 为最佳疗法，此时即为问题二的情况。相反，若病人对费用敏感(取 r=1 时)，疗法 1 的性价比函数 U 最大，所以对于对费用较敏感的病人来说，应选择第 1 种疗法。问题三中给的是供应商向不发达国家提供药品的价格，而对于一些不发达国家，他们对疗法费用的敏感系数相对较高(假定为 1)，则应选择疗法 1。

此时，我们注意到：有相当数目的病人用药没有效果，他们的最佳停药时间应为 0，若这些病人的数据参与拟合模型，将大大降低模型的效率，故我们在拟合模型前将这部分数据舍去。我们用问题一中效果最好的模型，也就是模型 3 的支持向量机算法，来对采取疗法 1 的病人确定终止治疗时间和预测疗效，其结果如表 10(仅随机取 10 个病人)。

| PtID | 终止时间 | 预测疗效（第 48 周）<br>按前述方法处理后的值 |
|---|---|---|
| 35 | 13 | 10.6 |
| 48 | 19 | 7.9 |
| 56 | 10 | 2.4 |
| 61 | 17 | 15.8 |
| 71 | 16 | 7.6 |
| 118 | 18 | 8.5 |
| 132 | 13 | 9.7 |
| 157 | 25 | 7.3 |
| 160 | 14 | 14.6 |
| 167 | 15 | 4.2 |

表 10

由于每个病人的测试样本过小，导致预测的误差较大。

表 10 中预测疗效的数值整体比表 9 中的小，从一个侧面说明了疗法 4 的疗效好于疗法 1。

## 6. 灵敏度分析

在求解问题三时，考虑了费用的影响时，我们仅考虑了 r=1 时的情况。但对于不同的人来讲，r 值是不同的，我们对 r=0.5，r=2 的情况进行讨论，研究费用敏感指数 r 的变化对性价比函数 U 的影响。如下表所示：

| U / 疗法 | r=0 | r=0.5 | r=1 | r=2 |
|---|---|---|---|---|
| 1 | 0.171887 | 0.155301 | 0.140316 | 0.114543 |
| 2 | 0.301678 | 0.162418 | 0.0874428 | 0.0253457 |
| 3 | 0.246551 | 0.157515 | 0.100633 | 0.0410746 |
| 4 | 0.380058 | 0.198931 | 0.104125 | 0.0285275 |

表 11

通过对表中数据进行分析可以得出，当 r 变化时，对疗法 4 的影响最大，当 r=0 时，疗法 4 是最优的，但当 r=2 时，疗效 4 几乎是最差的了，因此疗效 4 相对于 r 的变化很敏感。而当 r=0 时最差的疗法 1，r=2 时却变为了最优的疗法，说明疗法 1 相对了 r 的变化很不敏感。

我们进一步对 r 的变化对性价比函数的影响作了深入的研究，得到了效应性价比函数 U 关于敏感指数 r 的变化曲线，如图 2 所示：

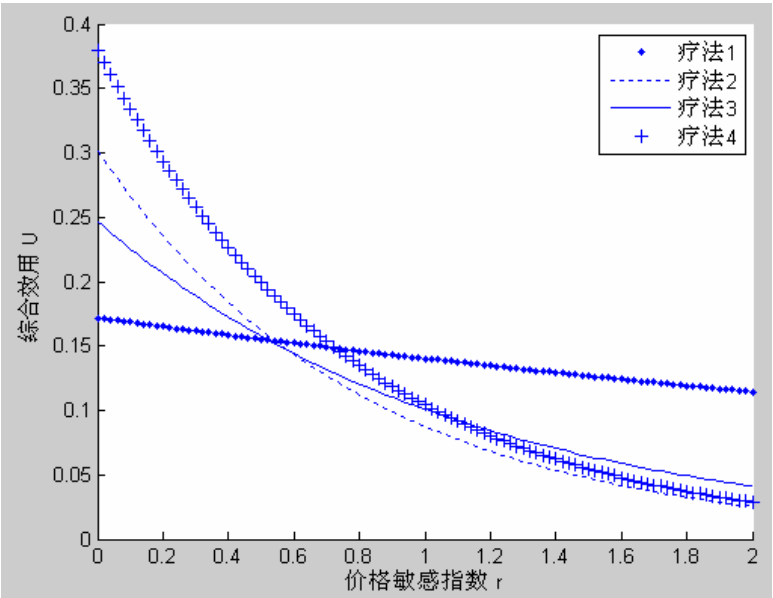

图 2

由图 2 我们可以很直观的看出，当费用敏感指数 r 增大时，对疗法 4 的性价比函数 U 减小的最快，说明疗法 4 的性价比函数相对于的 r 变化最敏感。相反，当 r 增大时，疗法 1 的性价比函数减小的速度最慢，即疗法 1 相对于 r 的变化最不敏感。

## 7．模型扩展

根据背景知识，目前的药物只能达到支持治疗的作用，而停药后病毒数往往会反弹，尤其是有些药物停药后就会对该药物产生抗药性。所以基于以上考虑，在没有过多毒副作用，且 CD4 的细胞数量能维持在一个相对可观的范围时，用药的时间应尽量延长。即用药的在很大程度上是续命的作用。

设 $f_j^i(t)$ 是第 $i$ 种疗法的第 $j$ 位病人第 t 周的 CD4 的数量，$g^i{}_j(t) = f^i{}_j(t)/f^i{}_j(0)$ 是对应的相对 CD4 细胞数量。对于不同的治疗方案，病人的平均相对细胞数量为 $h^i(t) = \sum_j g^i{}_j(t)/n_i$，（$n_i$ 为各种治疗方法对应的病人数量)，计算出不同疗法 $h^i(t)$ 的最大值 $\hat{h}^i$ 到达的时间 $t^i$。计算在 $t^i$ 后 $h^i(t)$ 的标准差 $s_i$，以反应反复的剧烈情况。

我们引入变量 $m_i$，定义 $h^i(t) \geq m_i * \hat{h}^i$ 的值为理想值。$s_i$ 在一定程度上反映了药效的稳定性和毒副作用的情况，所以 $m_i$ 与 $s_i$ 成反相关关系。一种可行的方案是，我们定义 $m_i = s_i \wedge (-0.1)$，此函数参数选择的优点是当 $s_i$ 在一个较大的范围内变化时，$m_i$ 的变化基本在 0.5 至 0.8 之间(例将 10 和 1000 分别代入计算得 $m_i$ 的值为 0.794328，0.501187)，比较符合人们的接受情况。

当 CD4 的细胞数量能维持在这一范围时尽量延长治疗时间，对应时间的最大值为停药时间。

对于每个病人来说，样本的数目比较少，这个计量经济学中的面板数据（PanelData）模型相似。但是我们已知的关于面板数据的模型都是线形的，而我们问题中的数据非线性特性非常明显，故我们没有采用这样的模型。

在生存分析中有一类 COX 半参数回归模型，是基于生存的概率分布的。若能找到一个合适的方式把我们数据中的目标转化成概率函数，则可以利用这个模型。遗憾的是我们没有找到这样的方法。

若对每个病人提供足够多的样本，我们可以利用时间序列分析的模型。那将大大增加我们模型的预测准确性。考虑到测试成本，我们建议减少被测病人的个数，但增加每个病人测试的次数。这样得到的数据已有利于我们建立更好的模型。

若能对每个病人提供更多的样本，则对于整个模型可以看成灰色系统。相应的我们可以利用已经比较成熟的灰色模型建立具有灰色指数规律的时间序列。如果数据允许，我们可以进一步建立有时变参数的时间序列进行分析预测。如设 $\{c_k\}$ 为一时间序列，$c_{k-1}c_{k-2}...c_{k-n}$ 已知，根据 $\{c_{k-1}c_{k-2}...c_{k-n}\}$ 可以预报 $c_k$ 的值 $\hat{c}_k$。

$$c_k = a_1(k)c_{k-1} + a_2(k)c_{k-2} + ... + a_n(k)c_{k-n} + b(k)$$

$a_1(k)a_2(k)...a_n(k)$ 为时变参数， $b(k)$ 为一维随机误差， $n$ 是模型阶数。

$\varphi(k)=[c_{k-1}c_{k-2}...c_{k-n}]^T$，

$\theta_k=[a_1(k)a_2(k)...a_n(k)]^T$,模型可以改写为： $\delta_k=\varphi(k)^T\theta(k)+b_k(2)$

然后采用随机补偿法将误差化到系统的时变参数上去，令

$$\beta(k)=\theta(k)+\frac{1}{\|\varphi(k)\|^2}\varphi(k)b_k$$

由于 $\varphi(k)^{\mathrm{T}}\beta(k)=\varphi(k)^T\theta(k)+\varphi(k)^T\frac{1}{\|\varphi(k)\|^2}\varphi(k)b_k=\varphi(k)^T\theta(k)+b_k$

所以有 $c_k=\varphi(k)^T\beta(k)$ 式中， $\beta(k)=[\beta_1(k)\beta_2(k)...\beta_n(k)]^T$ 是随机时变参数，$\{\beta(k)\}$ 构成一个时间序列，相对时间序列 $\{c_k\}$， $\{\beta(k)\}$ 是第二层时间序列。

应用拉普拉斯条件极值法获得时间序列 $\{\beta(k)\}$ 的递推公式，确定其部分估计值，然后依据这些数据判断时间序列 $\{\beta(k)\}$ 的平稳性，再确定用多维平稳随机模型或 AR 模型进行预测。

然后利用预测的数据作接下来的分析。

# 8．模型评价

## 8.1 优点：

1）基本求解中我们合理假设，简化关系。首先定义了的平均相对误差，用以衡量模型优劣程度。分别基于正态分布，多项式回归预测，支持向量机预测建立了三个模型，使用开放源代码的 Matlab 工具包 LibSVM 来执行支持向量机算法，得到的表达式从拟合度和稳定性两方面考虑都较为令人满意。

2 )在问题一，二的求解过程中，首先对数据进行了分析和处理，排除了相应的缺失数据和不匹配数据，删除了无效数据。通过Spearman 秩相关分析，研究了数据之间的相关关系。

3）在问题二中，我们给出了评价 4 种疗法优劣程度的效用函数，并利用问题一的支持向量机算法解决了最佳终止治疗时间的问题，虽然预测有偏差，但反映了最佳时间选择的一些规律；在问题 3 中引入价格敏感指数，来区别对待不同人对治疗费用的敏感程度的差异，提出了合理的性价比函数 U 。

4）在进一步的模型扩展中中我们放松假设，逼近现实，对停药时间和预测方法进行了进一步讨论。

### 8.2 缺点：

1 ） 由于每个病人的测试样本少，所求得的误差较大；
2 ） 使用支持向量机建模时，没有用到有病情反复现象的数据。

## 9．参考文献

[1] 孙山泽，非参数统计讲义，北京：北京大学出版社，2000 年
[2] 陈希孺，非参数统计，上海：上海科学技术出版社，1988 年
[3] Vapic P.，统计学习理论（中文版），北京：电子工业出版社，2000 年
[4] LibSVM，http://www.csie.ntu.edu.tw/~cjlin/libsvm/，2006-9-15
[5] 王开发，梁正东，免疫系统中 HIV 感染模型的动力学分析
http:/ /wanfang.calis.edu.cn/wf/~kjqk/myxzz/myxz2000/0002pdf/000222.pdf，2006-9-16
[6] 数学模型(第三版)，姜启源，北京：高等教育出版社，2003