Statistical and Application 统计学与应用, 2015, 4(4), 252-261
Published Online December 2015 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2015.44028

Hans汉斯

The Joint Model of Longitudinal and Survival Data

—Based on Machine Learning Methods

Zheng Wen

School of Mathematics, Yunnan Normal University, Kunming Yunnan
Email: yiyoubanren@163.com

Received: Dec. 2nd, 2015; accepted: Dec. 20th, 2015; published: Dec. 23rd, 2015

Abstract

In this paper, machine learning methods for longitudinal data and survival data modeling, replace the longitudinal sub-model linear random effects model; survival sub-model still uses Cox proportional hazards model. Compared with the traditional method, the residuals plots of survival sub-model diagnose modeling methods in line with theoretical results and the residuals of the longitudinal sub models are more dispersed than the linear mixed model.

Keywords

Joint Model, Machine Learning, Martingale Residuals, Cox-Snell Residuals

纵向数据与生存数据的联合模型

—基于机器学习方法

温　征

云南师范大学数学学院，云南　昆明
Email: yiyoubanren@163.com

收稿日期：2015年12月2日；录用日期：2015年12月20日；发布日期：2015年12月23日

文章引用: 温征. 纵向数据与生存数据的联合模型[J]. 统计学与应用, 2015, 4(4): 252-261.
http://dx.doi.org/10.12677/sa.2015.44028

摘　要

本文运用机器学习方法对纵向数据与生存数据建模，以机器学习方法代替纵向子模型中的线性随机效应模型；生存子模型仍运用Cox比例危险模型。与传统的建模方法做对比，此建模方法的生存子模型残差图诊断符合理论结果，纵向子模型的残差要比线性混合模型分散。

关键词

联合模型，机器学习，殃残差，Cox-Snell残差

1. 引言

在医学临床实验中经常收集到单个观测对象的多个指标的多次测量结果即纵向数据和时间事件数据。对于这类测量结果的研究经常被分开分析，但是在某种场合，感兴趣的却是这纵向结果和生存结果的相关结构。在过去的二十年中，这种数据的统计分析引起了统计学家的关注。纵向数据和时间事件数据的联合模型也随之发展，并广泛应用于医学中随访数据的分析。Rizopoulos [1]等给出了这类测量结果研究的综述。

纵向数据和时间事件数据的联合模型分为两个部分：纵向子模型和生存子模型。纵向子模型是线性随机效应模型，Diggle [2]等给出了这类模型研究的综述。在随机效应服从正态分布的假设下，得到了一些估计方法，例如 Laird 和 Ware [5]；Breslow 和 Clayton [3]；Hedeker 和 Gibbons [6]。正态性的假设在数学上带来了方便，然而 Fattinger [4]等指出：这一假定并不一定总是成立的。当随机效应的分布假设不成立时，估计量的效率可能受到损害。已经有几位学者研究了没有正态性假设时的估计方法，如 Fattinger [4]等，Magder [7]等，Kleinman [8]等以及 Tao [9]等。生存子模型通常是 Cox 比例危险模型。Cox 模型由 Cox [10]于 1972 年提出，处理与时间无关的协变量。Andersen 和 Gill [11]于 1982 年给出了扩展的 Cox 模型，Fleming 和 Harrington [12]于 1991 年对扩展的 Cox 模型做了详细的描述。因此扩展的 Cox 模型也叫做 Andersen-Gill 模型。

纵向数据和生存数据的联合模型的标准模型如下：

纵向子模型：

$$\begin{cases} y_i(t) = m_i(t) + \xi_i(t) \\ m_i(t) = x_i^{\mathrm{T}}(t)\beta + z_i^{\mathrm{T}}(t)b_i \\ b_i \sim N(0, D), \xi_i(t) \sim N(0, \sigma^2) \end{cases} \tag{1}$$

其中 $y_i(t)$ 表示 t 时刻个体的标识物的测量结果，$m_i(t)$ 表示标识物的真实水平，$\xi_i(t)$ 表示误差项，其服从正态分布 $N(0, D)$。x_i^{T} 表示固定效应 β 的设计向量。Z_i^{T} 表示随机效应 b_i 的随机效应。假设 b_i 服从 $N(0, D)$

生存子模型：

标准比例危险模型：假设协变量与时间无关。

$$h_i(t) = h_0(t)\exp\{\gamma^{\mathrm{T}} w_i\},\ t > 0$$

扩展的比例危险模型：

$$h_i(t)=h_0(t)\exp\left\{\gamma^{\mathrm{T}}w_i+\alpha m_i(t)\right\},\ t>0 \tag{2}$$

式中 $h_0(t)$ 表示基准危险函数即标准比例危险模型中当协变量取值为 0 或者说 $\gamma^{\mathrm{T}}w_i=0$ 时的危险函数。w_i 表示基准协变量向量，γ^{T} 为其相应的回归系数向量。$m_i(t)$ 表示依时间变化的标识物的当前值，含义与纵向子模型中的 $m_i(t)$ 相同。α 为纵向结果(纵向标识物)回归系数或者说代表纵向结果与生存结果的关联性，其大小表示纵向结果与生存子模型中危险函数的关联性的强弱。

然而，正如上述所言纵向子模型中的混合随机效应模型需要对数据做各种各样的假设，这些假设，只是数学上方便但不一定满足，因此结果会产生偏倚。近年来机器学习的方法即算法模型被广泛应用，正如吴喜之[13] [14]所说在处理巨大的数据集上，在对付被称为维数诅咒的巨大变量书模式在无法假定总体分布的情况下，在面对众多竞争模型方面，算法建模较经典模型有很多不可比拟的优越性。

2. 建立模型

2.1. 数据来源及 R 软件介绍

本文用到的数据集是 R 软件 JM 程序包自带的 AIDS 数据集。R 软件是一个开放的统计编程环境，是一种语言，是一套完整的数据处理、计算和制图软件系统。

2.2. 建立模型

1) 传统方法

传统方法也就是纵向子模型在正态性假定条件下的线性混合效应模型，公式(1)。生存子模型是 Cox 模型，公式(2)。

从表 1 中可以看出，除了纵向过程中 obstime*drugddI 项的系数在显著性水平 0.05 下，不显著外，其他检验都比较显著。

传统的联合模型的纵向子模型是在正态分布的假设下，才能成立，正如 Fattinger [4]等指出：这一假定并不一定总是成立的。当随机效应的分布假设不成立时，估计量的效率可能受到损害，产生偏倚。现检验残差的正态性。

从图 1 和表 2 即残差的 QQ 和 Shapiro 正态性检验中都可以看出残差不服从正态分布，即纵向子模型的假设不成立。

2) 机器学习方法

本文在纵向子模型中，用机器学习方法来代替线性混合效应模型做回归分析。机器学习方法主要用决策树回归、Bagging 回归、随机森林回归、最邻近方法回归、支持向量机回归五种方法分别对纵向数据做回归分析。因变量为 CD4，其他变量为自变量，比较其标准均方误差选择最合适的方法用于联合建模。标准均方误差公式为：

$$NMSE=\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

生存子模型不变。但生存子模型中的 $m_i(t)$ 用纵向子模型中的回归结果来代替。R 代码见附录。

从表 3 中可以看到机器学习的五种方法的标准均方误中最小的是最邻近方法回归。因此它是拟合 aids 数据最好的方法。以此方法对 aids 数据建立纵向子模型。

以 aids 数据中变量 drug 为基准协变量，纵向子模型中回归树所得结果作为 $m_i(t)$，做 Cox 回归。

表 4 中的 drugddI 系数比表 1 中的 drugddI 系数明显小很多，表 4 中 $m_i(t)$ 的系数绝对值要比表 1 中 $m_i(t)$ 系数的绝对值要小；从表 5 中可得，模型的似然比检验的 p 值为 0；Wald 检验的 p 值为 0；Score

Table 1. The part results of joint model
表 1. 联合模型部分结果

	Longitudinal Process			Event Process	
变量	Intercept	obstime	Obstime∗drugddI	drugddI	Assoct ($m_i(t)$)
系数值	7.2203	−0.1917	0.0116	0.3348	−0.2875
P 值	0.0001	0.0001	0.7014	0.0324	0.0001

Table 2. The result of Shapiro test
表 2. Shapiro 正态性检验结果

Shapiro-Wilk normality test		
data	W statistics value	P-value
reslmefit	0.94619	2.2×10^{-16}

Table 3. The results of the regressions of longitudinal submodel
表 3. 纵向子模型的回归方法结果

方法	决策树	Bagging	随机森林	最邻近方法	支持向量机
NMSE	0.1641386	0.2963151	0.7226739	0.1559414	0.2361392

Table 4. The results
表 4. 结果

机器学习与 Cox 比例危险模型的联合模型		
变量	drugddI	$m_i(t)$
系数值	0.23238	−0.20830
P 值	0.0187	2×10^{-16}

Table 5. The tests of Cox model
表 5. Cox 模型检验

检验	Likelihood ratio test	Wald test	Score (logrank) test
统计量值	208.9	144.8	165.1
P 值	0	0	0

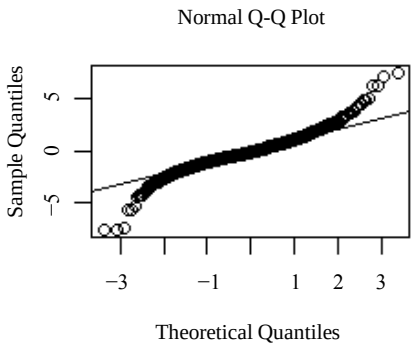

Figure 1. Normal QQ plot of residuals
图 1. 残差的正态 QQ 图

检验的 P 值为 0。除了 R^2 较小外，三大检验比较显著，似乎拟合得还可以。

3. 模型诊断

3.1. 纵向子模型的残差对比

从图 2 中可以看出：左边 kknn 法得到的残差的趋势是在大概 7 位置之前残差为负，之后为正；右边线性混合效应模型的残差趋势是在大概 7 位置之前残差为负，之后为正。在大概 7 的位置趋势线分成两段，kknn 法的趋势线的斜率要比线性混合效应模型的趋势线的斜率大。

从图 3 中看出：左图，最邻近方法回归的残差散点集中在水平线−6 到水平线 10 之间，主要集中在−6 到 6 之间；右图，线性混合效应残差的散点图范围在−8 到 8 之间，主要集中在−5 到 5 之间。最邻近方法要比线性混合效应回归的残差分散。

3.2. 生存子模型的残差对比

生存模型：Cox 回归模型的残差与传统的残差不同。本文用到的是殃残差和 Cox-Snell 残差。传统的联合模型可以直接提取 Cox-Snell 残差。而最邻近方法的生存子模型不能够直接提取 Cox-Snell 残差，但能提取殃残差经以下公式可转换成 Cox-Snell 残差。转换公式：

殃残差公式：

$$r_i^{tm}(t) = N_i(t) - \int_0^t R_i(s)\hat{h}_0(s)\exp\{\hat{\gamma}w_i + \hat{\alpha}m_i(t)\}\mathrm{d}s \tag{3}$$

$$= \delta_i - \hat{H}(\tau_i)\mathrm{e}^{\hat{\beta}'X_i} \tag{4}$$

Cox-Snell 残差公式：

$$r_i^{tcs} = \int_0^{T_i} R_i(s)\hat{h}_0(s)\exp\{\hat{\gamma}w_i + \hat{\alpha}m_i(t)\}\mathrm{d}s \tag{5}$$

$$= N_i(T_i) - r_i^{tm}(T_i)$$

$$= \delta_i - r_i^{tm} \tag{6}$$

图 4 中，左上部分是最邻近方法联合模型的 Cox-Snell 残差的 KM 估计，右上是最邻近方法 Cox-Snell 残差对 drug 的 KM 估计；左下是线性混合效应联合模型的 Cox-Snell 残差的 KM 估计，右上是线性混合效应联合模型的 Cox-Snell 残差对 drug 的 KM 估计。虚线是其相应的 95%逐点的置信区间，灰色实体线是参数为 1 的指数分布函数。

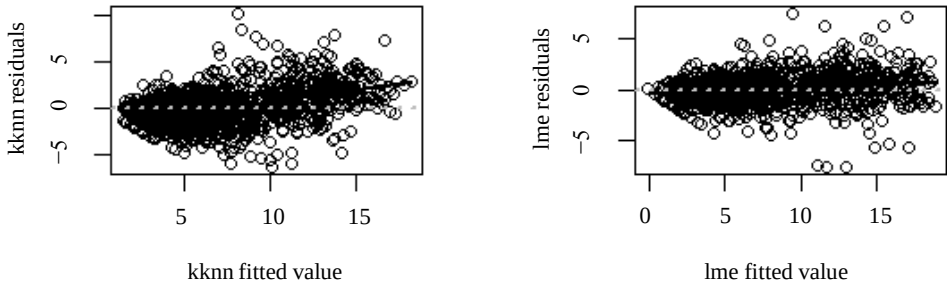

Figure 2. The plot of residuals and the fitted values of the linear mixed-effect model and kknn method
图 2. 线性随机效应回归和最邻近方法回归的残差与拟合值的图像

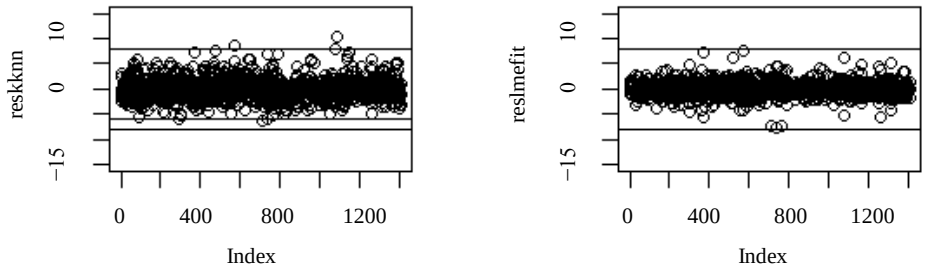

Figure 3. The residuals plot of the linear mixed-effect model and kknn method
图 3. 最邻近方法和线性混合效应的残差图

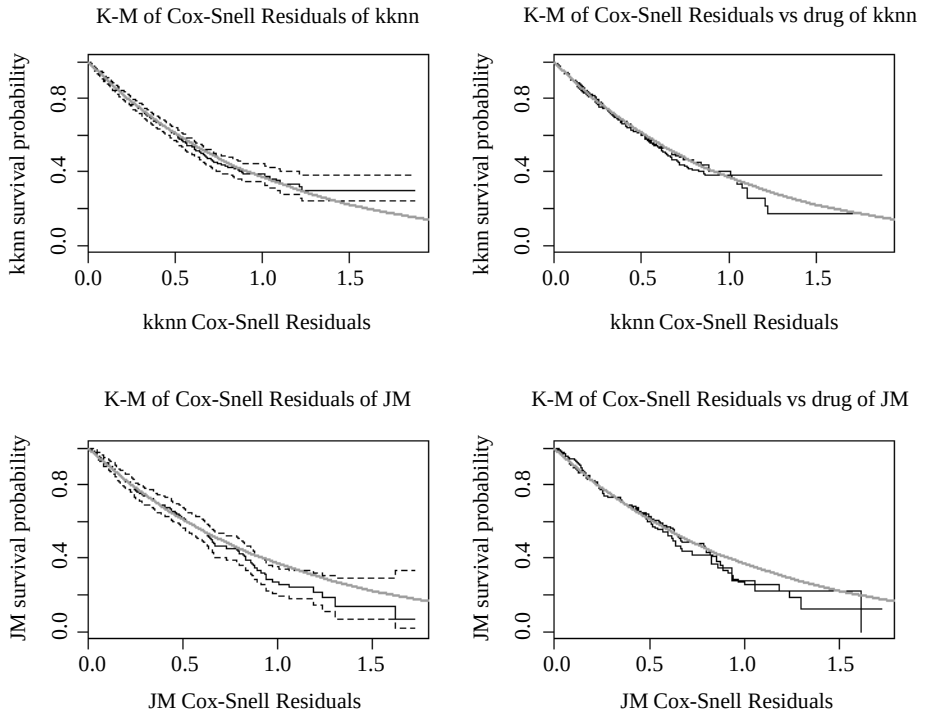

Figure 4. The comparison plot of two methods
图 4. 两种方法的对比图

当模型拟合的数据比较好时，Cox-snell 残差的生存函数的估计应该围绕参数为 1 的指数分布函数图像变动。残差图 3 中，对比左上和左下两图，可以看出左上图大概在横坐标为 1 的位置开始灰色与 Cox-snell 残差的 KM 估计图像不在相交，而左下图从大概横坐标为 0.8 的位置开始就不再相交。右上和右下的图像对比，最邻近方法联合模型不同药物治疗效果的 Cox-snell 残差的 KM 估计图像比线性混合效应联合模型不同药物治疗效果的 Cox-snell 残差的 KM 估计图像更接近理想状态。因此最邻近方法联合模型生存部分模型拟合的很好。

4. 结论

这种纵向数据和生存数据的联合模型的新方法，即纵向子模型用机器学习方法做回归，生存子模型不变，依据 Dimitris 的诊断方法，拟合出的生存模型比传统的联合模型的生存子模型好，其残差图像更符合理论结果。此外机器学习中的纵向子模型不用考虑各种假设分布，且能充分利用观测变量和观测信息，但就本文而言其残差图比线性混合效应模型的残差图更分散。传统方法中纵向子模型为线性混合效

应模型，在建立模型之后需要检验模型的假设条件是否成立。当假设条件成立时，线性混合效应模型模型的标准均方误差要比机器学习方法小很多，此外传统方法比较成熟，因此使用传统方法较好；当不成立时，用机器学习方法建立的联合模型较好，不用考虑违背假设时产生偏倚情况。该文章是对一种新的纵向数据和生存数据联合建模的方法的探讨，仍存在很多不足之处，需要继续深入研究。

参考文献 (References)

[1] Rizopoulos, D. (2012) Joint Models for Longitudinal and Time-to-Evwnt Data with Applications in R. Chapman &Hall/CRC Biostatistics Series, 51-155.

[2] Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002) Analysis of Longitudinal Data. 2nd Edition, Oxford University Press, Oxford.

[3] Breslow, N.E. and Clayton, D.G. (1980) Approximate Inference for Stochastic Process. Academic Press, London.

[4] Fattinger, K.E., Sheiner, L.B. and Verotta, D. (1995) A New Method to Explore the Distribution of Inter Individual Random Effects in Non-Linear Mixed Effects Model. *Biometrics*, **51**, 1236-1251. http://dx.doi.org/10.2307/2533256

[5] Laird, N. and Ware, J.H. (1982) Random-Effects Models for Longitudinal Data. *Biometrics*, **38**, 963-974. http://dx.doi.org/10.2307/2529876

[6] Hedeker, D. and Gibbons, R.D. (1994) A Random Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics*, **50**, 933-953. http://dx.doi.org/10.2307/2533433

[7] Magder, L.S. and Zeger, S.L. (1996) A Smooth Nonparametric Estimate of a Mixing Distriution Using Mixtures of Gaussians. *Journal of the American Statistical Association*, **91**, 1141-1151. http://dx.doi.org/10.1080/01621459.1996.10476984

[8] Kleinman, K.P. and Ibrahim, J.G. (1998) A Semiparametric Bayesian Approach to the Random Effects Model. *Biometrics*, 921-938.

[9] Tao, H., *et al*. (1999) An Estimation Method for the Semiparametric Mixed Effects Model. *Biometrics*, **55**, 102-110. http://dx.doi.org/10.1111/j.0006-341X.1999.00102.x

[10] Cox, D. (1972) Regression Models and Life-Tables (with Discussion). *Journal of the Royal Statistical Society*, *Series B*, 187-220.

[11] Andersen, P. and Gill, R. (1982) Cox's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics*, **10**, 1100-1120. http://dx.doi.org/10.1214/aos/1176345976

[12] Fleming, T.R. and Harrington, D.P. (1991) Counting Processes and Survival Analysis. Wiley, New York.

[13] 吴喜之. 统计学: 从数据到结论[M]. 第四版, 北京: 中国统计出版社, 2014.

[14] 吴喜之. 复杂数据统计方法——基于 R 的应用[M]. 第二版, 北京: 中国人民大学出版社, 2013.

附　录

```
library(JM)
w=aids[,-c(10,11,12)]
attach(w)
########################传统方法
##########纵向子模型
lmefit=lme(CD4~obstime +obstime:drug,random=~obstime| patient,data=w)
(NMSE=mean((w$CD4-predict (lmefit,data=w))^2)/mean((w$CD4-mean(w $CD4))^2))
summary(lmefit)
##########生存子模型
coxfit=coxph(Surv (Time,death)~drug,data=aids.id,x=T)
summary(coxfit)
##########联合模型
jointfit=jointModel (lmefit,coxfit,timeVar="obstime",method ="piecewise-PH-aGH")
summary(jointfit)
##########纵向子模型的残差 QQ 图及 Shapiro  检验
reslmefit=residuals(lmefit)
qqnorm(reslmefit)
qqline(reslmefit)
shapiro.test(reslmefit)
########################机器学习方法
##########分类树回归
library(rpart.plot)
cf1=rpart(CD4~.,data=w)
(NMSE=mean((w$CD4-predict(cf1,data=w))^2)/mean((w$CD4-mean(w$CD4))^2))
##########Bagging 回归
library(ipred)
set.seed(110)
cf2=bagging(CD4~.,data=w,coob=T,control=rpart.control(xval=10))
(NMSE=mean((w$CD4-predict(cf2,data=w))^2)/mean((w$CD4-mean(w$CD4))^2))
##########随机森林回归
library(randomForest)
set.seed(110)
cf3=randomForest(CD4~.,data=w[,-1],importance=T,proximity=T)
(NMSE=mean((w$CD4-predict(cf3,data=w[,-1]))^2)/mean((w$CD4-mean(w$CD4))^2))
##########最邻近方法
library(kknn)
set.seed(110)
cf4=kknn(CD4~.,train=w,test=w)
```

```
cf4fit=cf4$fit
(NMSE=mean((w$CD4-cf4fit)^2)/mean((w$CD4-mean(w$CD4))^2))
##########支持向量机
library(rminer)
set.seed(110)
cf5=fit(CD4~.,w,model="svm")
y=predict(cf5,w)
(NMSE=mean((w$CD4-y)^2)/mean((w$CD4-mean(w$CD4))^2))
##########生存子模型
library(survival)
sf=coxph(Surv(Time,death)~drug+cf4fit,w)
summary(sf)
##################纵向部分残差对比
par(mfrow=c(1,2))
#######机器学习方法 kknn
reskknn=w$CD4-cf4fit
plotResid=function(x,y,col.loess="black",...){
plot(x,y,...)
lines(lowess(x,y),col=col.loess,lwd=2)
abline(h=0,lty=3,col="grey",lwd=2)}
plotResid(cf4fit,reskknn,xlab="kknn fitted value",ylab="kknn residuals")
#######线性混合效应模型
fitvalue=fitted(lmefit)
plotResid(fitvalue,reslmefit,xlab="lme fitted value",ylab="lme residuals")
plot(reskknn,ylim=c(-15,15))
abline(h=8);abline(h=-8);abline(h=-6)
plot(reslmefit,ylim=c(-15,15))
abline(h=8);abline(h=-8)
##################生存部分残差比较
par(mfrow=c(2,2))
#########kknn 方法
es=residuals(sf,"martingale",collapes=T)
es1=death-es#CoxSnell 残差
aa=Surv(es1,death)
sfit=survfit(aa~1)
plot(sfit,mark.time=F,xlab="kknn Cox-Snell Residuals",ylab="kknn survival probability",main="K-M of Cox-Snell Residuals of kknn")
curve(exp(-x),from=0,to=max(w$Time),add=T,col="grey62",lwd=2)
sfit1=survfit(aa~drug)
plot(sfit1,mark.time=F,xlab="kknn Cox-Snell Residuals",ylab="kknn survival probability",main="K-M of
```

```
Cox-Snell Residuals vs drug of kknn")
curve(exp(-x),from=0,to=max(w$Time),add=T,col="grey62",lwd=2)
##########线性混合效应
resCS=residuals(jointfit,process="Event",type= "CoxSnell")
sfit3=survfit(Surv(resCS,death) ~ 1, data = aids.id)
plot(sfit3,mark.time=F,xlab="JM Cox-Snell Residuals",ylab="JM survival probability",main="K-M of Cox-Snell Residuals of JM")
curve(exp(-x),from=0,to=max(w$Time),add=T,col="grey62",lwd=2)
sfit4=survfit(Surv(resCS,death) ~ drug, data = aids.id)
plot(sfit4,mark.time=F,xlab="JM Cox-Snell Residuals",ylab="JM survival probability",main="K-M of Cox-Snell Residuals vs drug of JM")
curve(exp(-x),from=0,to=max(w$Time),add=T,col="grey62",lwd=2)
```