**缺失值处理实例分析**

本例中的数据是纽约市家庭的社会指标电话调查数据，每两年由哥伦比亚大学社会工作学院组织进行的。请进行如下步骤的数据分析工作：

**1.读入数据及数据准备**

（1）将数据siswave.csv读入R后命名为wave，使用attach(wave)函数激活该数据集。

（2）生成名为earnings的新变量列（要求：单位是***$1000***，原始数据单位是$），表示家庭总收入，它是由rearn(响应者收入)和tearn（响应者配偶收入）两部分加起来生成的。注意要将rearn和tearn中的负数分配为NA后再求和。统计earnings变量的缺失百分比。

（3）生成名为white的新变量列，若race of respondent（1=white, 2=black, 3=hispanic(nonblack), 4=other）是1则white=1，否则white均为0。考虑如果race变量有缺失的情况，若缺失则white也为0。（这个步骤就是一个训练数据处理的步骤，如果原数据中有white覆盖掉即可）

（4）生成名为male的新变量列，男性male=1，女性male=0。依据是wave中的变量列sex: male是1, female是2。

（5）生成名为over65的新变量列，年龄超过65则over65=1，否则over65=0。依据是wave中的年龄变量r\_age。

（6）将变量immig中的缺失值均用0插补。immig=0表示是美国公民, immig=1 则不是美国公民。

（7）educ\_r表示不同的教育水平。如果是NA，用2.5替换。1至4表示不同的教育水平，依次增高。

（8）分别基于变量ssi，welfare，charity，生成名为any.ssi，any.welfare，any.charity的变量，将正值赋值为1，其他负值或缺失值都赋值为0。

ssi: supplemental security Income for entire family

welfare: public assistance for entire family

charity: income received from charity for entire family

-9: 响应者拒绝回答是否有此来源收入；

-5：响应者有此项收入但没提供收入数目

**2.缺失模式探索（Exploration of Missing pattern）**

（1）基于原始数据wave，生成一个新的数据框，命名为sis.sm，它依次包括sex,race,edu\_r,r\_age,earnings,police。给这些不同列合适的类型，比如sex,race,police都设置成因子型。edu\_r设置为有序的因子。

（2）判断earnings是否是MCAR

判断方法可以简单的看一下不同种族race是黑人和白人情况下earnings的观测数、缺失观测数和缺失频率。如果他们的缺失频率不相等，那么说明种族race会影响earnings是否缺失。可判断它就不是MCAR。

（3）使用mice包里的md.pattern()函数查看数据的缺失模式。

可视化它的观测值和变量列的缺失状况。

（4）最简单的方法是删除有缺失的观测。这里，如果观测中的缺失变量个数百分比超过30%，那么删除这个缺失观测。统计删除了多少观测，保留多少观测。

**3.简单随机插补（simple random imputation）**

简单随机插补是从存在缺失的变量的已有观测值中随机抽取一个值来插补缺失值。本小题的任务是定义一个名为random.impute的R函数，功能是对收入earnings进行随机插补，返回一个完整的数据。

>random.impute<-function(a)

{……#请填写中间步骤

}

>earnings.impute<-random.impute(earnings) #使用该R函数对earnings进行简单随机插补，生成插补后的完整数据earnings.impute。

**4.回归插补（Using regression predictions to perform deterministic imputation）**

（1）生成新变量earnings.top，将收入earnings超过$100,000的修改为$100,000。目的是避免收入特别高的观测值影响回归结果，有些收入甚至上百万美元。

（2）生成新变量workhrs.top，将工作小时数workhrs中每周超过40小时的设置为40小时。虽然有些响应者确实工作时间会超过40小时，这样简单的变换可能提高回归模型的预测性能，避免极端值的影响。

（3）生成一个数据框名为SIS，包含变量列earnings，earnings.top，male，over65，white，immig，educ\_r，workmos，workhrs.top，any.ssi，any.welfare，any.charity.

（4）以earnings中的***非零且非缺失观测值***为被解释变量，建立一个回归模型，自变量是male，over65，white，immig，educ\_r，workmos，workhrs.top，any.ssi，any.welfare，any.charity.回归模型的结果命名给lm.impute.1。

>lm.impute.1<-lm(earnings~….,data=SIS, subset=… ) #自己补充这部分程序

（5）将earnings中的缺失值使用模型lm.impute1的预测值进行插补。

（6）为预测中低收入的群体，对收入earnings.top进行平方根变换后当作被解释变量建立回归模型，回归模型结果命名为lm.impute.2.sqrt。

>lm.impute.2.sqrt<-lm(I(sqrt(earnings.top))~….) #自己补充这部分程序

>predict(…)

（7）将earnings.top中的缺失值使用模型lm.impute.2.sqrt的预测值进行插补，对比插补前earnings.top的直方图（去掉0收入值）和插补值的直方图（只有两百多个缺失值的插补值）。

**5.随机回归插补（Random regression imputation）**

通过在回归模型中增加误差可以将不确定性添加到插补值。插补方法是从正态分布（使用回归模型lm.impute.2.sqrt的估计结果）中随机抽样作为插补值。

提示：使用R函数：rnorm()函数进行抽样再插补。

**6. 两阶段插补**

该社会指数调查数据中，重点是插补收入earnings为正值的缺失值。若workhrs=0且workmos=0，则插补收入为0即可。若二者之一非零，插补earnings一个连续正值。这里，假设workhrs和workmos两个变量的信息是未知的，故而我们无法直接确定零收入的插补值。所以这里使用logisitic回归判定首先判定收入是否为正值，然后再进行随机回归插补。分如下两个步骤进行插补：

1. 插补收入是否为正的indicator variable，

。

插补方法是使用其他所有的自变量（male，over65，white，immig，educ\_r，workmos，workhrs.top，any.ssi，any.welfare，any.charity）建立的Logisitic回归模型

。

>glm.indicator<-glm(I(earnings>0)~….,data=SIS,family=binomial(link=logit))

#自己补充这部分logisitic回归模型的估计

对缺失观测，使用模型glm.indicator预测earnings为正的概率。然后从二项分布Binomial(1, )随机一个样本作为的插补值。

第二，基于回归模型lm.impute.2.sqrt，使用随机回归插补的方法，为上一步中收入判定为正(即)的缺失值插补一个连续正值。

**7. 迭代回归插补(iterative regression imputation)**

如果有多个变量有缺失，可通过将单变量的插补方法推广到多变量。如果缺失数据矩阵Y，各变量列分别是Y(1),Y(2)…Y(k)，完整的观测预测变量是X。

1. 首先将Y的各列变量使用粗糙的方法进行插补，比如简单随机插补。
2. 在给定Y(2)…Y(k)和X条件下，建立以Y(1)为因变量的回归模型，使用随机回归插补的方法插补Y(1)；
3. 与上步方法类似，在给定Y(1)，Y(3)…Y(k)和X条件下插补Y(2)，其中Y(1)使用上一步中更新后的插补值

……

1. 依次方法迭代插补，直到每个变量的插补值近似收敛。

将wave中的变量列interest中的负数设置为NA，使用迭代回归同时插补两列变量interest和earnings。回归模型中包括其他所有的完整变量（male，over65，white， immig， educ\_r， workmos，workhrs.top，any.ssi，any.welfare，any.charity）作为预测变量X。

检查你的插补结果，评价这种方法做出的插补结果的好坏。比如可以通过观测数据和插补后数据的直方图或散点图来评价。