1.Introduction

本文通过对给定的数据集进行预处理,观察数据后尝试多种回归模型，并用交叉验证的方法选出在测试集上预测均方误最小的模型-半参数回归模型，并对该模型估计的参数部分使用重抽样的方法计算方差，进行统计推断，给出估计参数的置信区间。

2.Analysis

2.1 Data

```
##       Y    X1 X2    X3
## 1 4.103 2.177 NA 0.808
## 2 4.179 2.964  0 0.541
## 3 3.300 2.425  0 0.724
## 4 2.199 1.656  0 0.048
## 5 4.217 2.631  1 0.295
## 6 5.203 2.683 NA 0.915
```

数据共有184行，4列，第一列Y为被解释变量，后三列为解释变量，其中Y,X1,X3为连续变量，X2为0-1变量且存在缺失值。缺失数据占X2总数据的11.4%。对数据进行正态性检验，其中Y,X1可视为正态分布，X3不可视为正态分布.

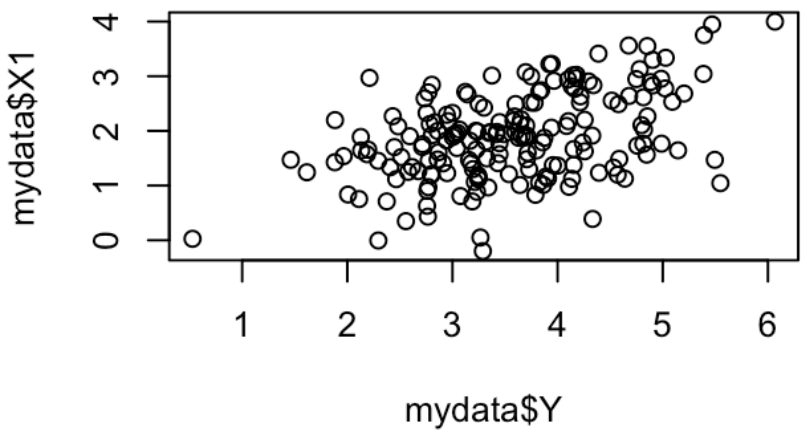
在数据预处理中离散型数据缺失值可以使用中位数，边际分布等方式进行填充，本文分别采用中位数，边际分布（伯努利分布），以及Y对X2进行逻辑回归估计三种方式填充缺失值(附录)，填补后进行多元线性回归分析，最终选择使得在测试集中预测均方误最小的中位数进行填充。

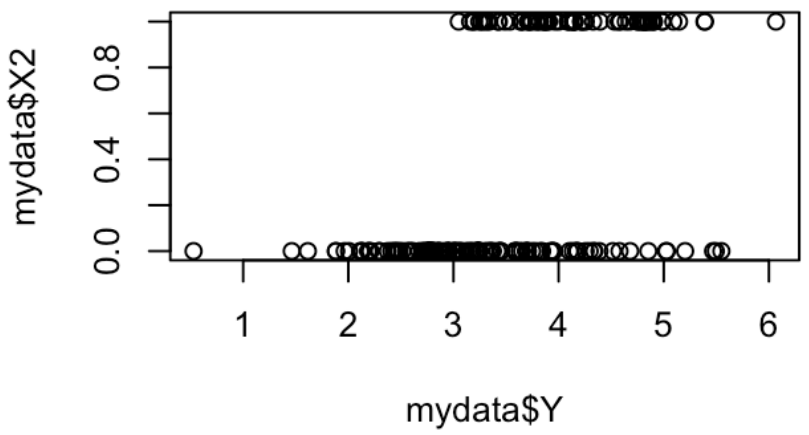
2.2 Analysis

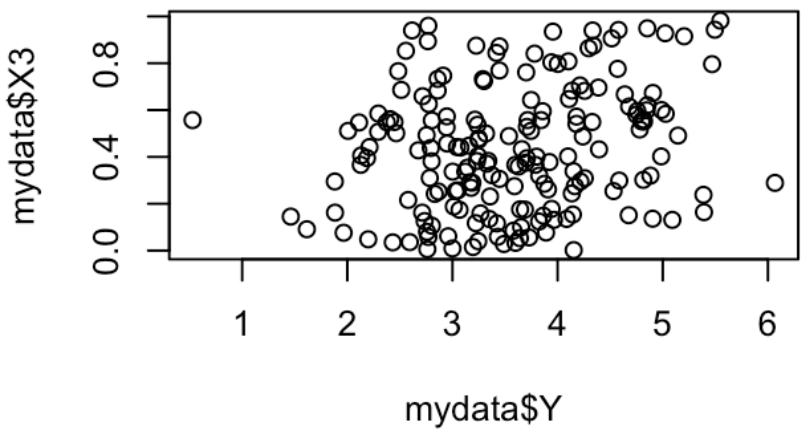
由于Y为连续型变量，考虑回归分析，回归分析包括简单线性回归，多元线性回归，多项式回归，对数线性回归，岭回归，lasso回归，半参数回归等，对数据进行初步分析，确定备选模型后，用交叉验证的方式选择使得预测均方误最小的模型。

观察数据的相关性，从散点图可看出中X1，X2与Y的线性关系较高，各解释变量之间相关性并不是非常高，故考虑先进行多元线性回归。

```
##             Y          X1          X2          X3
## Y  1.0000000  0.46573536  0.49992520  0.22247815
## X1 0.4657354  1.00000000 -0.03693411 -0.04148435
## X2 0.4999252 -0.03693411  1.00000000 -0.20194218
## X3 0.2224782 -0.04148435 -0.20194218  1.00000000
```








```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68986 -0.36802  0.01686  0.41323  2.19414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.49258    0.14450  10.329  < 2e-16 ***
## X1           0.58368    0.05411  10.787  < 2e-16 ***
## X2           1.13361    0.09107  12.448  < 2e-16 ***
## X3           1.27486    0.16706   7.631 1.31e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5815 on 180 degrees of freedom
## Multiple R-squared:  0.6107, Adjusted R-squared:  0.6042
## F-statistic: 94.11 on 3 and 180 DF,  p-value: < 2.2e-16
```

多元回归估计结果各项系数均显著，调整R方为0.6042，计算该各个变量的VIF值，VIF用于衡量变量直接的多重共线性，当VIF大于10时，视为存在多重共线性。本数据各个变量的VIF值均小于10，且变量间相关性不强，故不考虑岭回归与lasso回归

将备选模型分为四大类，每一类内中尝试不同组合，如多元回归中进行逐步回归，多项式回归时加多项式拟合，将四个大类内的模型，以R方为筛选标准，选出拟合最好的四个模型用于最后的交叉验证筛选(该步骤可在附录中查看)最后所选的四个模型如下：

1.多元线性回归

```
fit1 <- lm(Y~X1+X2+X3,data=mydata)
```

2.多项式回归

```
fit2 <- lm(Y~I(X1^2)+X2+X3,data=mydata)
```

3.对数变换回归,对(X3)取对数

```
fit3 <- lm(Y~X1+X2+log(X3),data=mydata)
```

4.半参数回归，将X1，X2视为参数部分，X3视为非参数部分

```
fit4 <- gam(Y~X1+X2+s(X3),data=mydata)
```

上述四个模型中的系数估计均显著，且R方大小接近。

交叉验证是指将数据集切分为训练集和测试集，在训练集中进行模型拟合，估计参数，在测试集中使用已经估计得到的参数进行预测，并计算预测值与实际值直接的误差，由于误差有正有负，加总时会存在正负相抵的情况，故在多次计算中可以采用绝对误差，或者是预测均方误对预测的准确率进行评估。本文通过交叉验证在上述四个模型中选择的最优模型为半参数回归模型，计算的预测均方误如下：

```
> c(pmse1,pmse2,pmse3,pmse4)
[1] 0.3565272 0.3550381 0.4001451 0.3351907
```

半参数回归模型为Y~b1*X1+b2*X2+h(X3),其中X1与X2视为模型的参数部分，h(X3)视为模型的非参数部分。其参数部分的系数估计为：

```
> b_hat
        mydata.X1 factor.mydata.X2.
        0.5638013         1.1751488
```

对半参回归模型得到估计系数后，使用bootstrap的方法对估计样本进行重抽样，得到估计方差，构造估计的置信区间。bootstrap即对现有的数据做重抽样，由于只有一个数据集，故一次估计中只能得到唯一的估计系数，但如果使用有放回的重抽样，就可以计算出一组估计参数，并得到估计方差，再利用渐进正态性，可以计算出估计参数的置信区间。

其参数部分的估计系数的置信区间为：

```
> rbind(b_n_lower, b_n_upper)
          mydata.X1 factor.mydata.X2.
b_n_lower 0.4350540          1.046402
b_n_upper 0.6925486          1.303896
```

3.Conclusion

本文通过尝试多种模型，最后选择预测均方误最小的半参回归模型Y~b1*X1+b2*X2+h(X3)。

其中参数部分系数的估计值为：

```
> b_hat
        mydata.X1 factor.mydata.X2.
        0.5638013         1.1751488
```

系数估计的置信区间为：

```
> rbind(b_n_lower, b_n_upper)
          mydata.X1 factor.mydata.X2.
b_n_lower 0.4350540          1.046402
b_n_upper 0.6925486          1.303896
```

4.Appendix

1.缺失值填补选择
将三种填补方式进行同一种回归模型估计（多元线性回归），计算其预测均方误差。
1.1边际分布（伯努利分布）填补
```
mydata[which(is.na(mydata$X2)),3] <- rbinom(m, 1, p)
```
由于按分布填充过程有随机性，故重复填充M次，没填充一次做一次多元线性回归，计算多次估计系数的平均值，再使用交叉验证，计算其PMSE
1.2中位数（0）填补
```
mydata[which(is.na(mydata$X2)),3] <- median(mydata[-which(is.na(mydata$X2)),3])
```
交叉验证计算PMSE
1.3y对x2逻辑回归填补
将Y与X2中未缺失的部分进行逻辑回归拟合，拟合后用Y预测X2的缺失值
```
mydata[which(is.na(mydata$X2)),3] <- X2_hat
```

2.四类模型的初筛选
###model1：
线性回归
对缺失数据进行插值，进行MLS估计
```
fit1 <- lm(Y~X1+X2+X3,data=mydata)
```
因缺失数据列的占比为11.4%较大，故考虑直接去掉数据X2,进行MLS估计
```
fit1 <- lm(Y~X1+X3,data=mydata)
fit1 <- lm(Y~X1,data=mydata)
```
###model2：
多项式回归，交叉项回归
```
fit2 <- lm(Y~I(X1^2)+X2+X3,data=mydata)#0.6599
fit2.1 <- lm(Y~X1+X2+I(X3^2),data=mydata)#系数不显著
fit2.2 <- lm(Y~X1+X2+X3:X2,data=mydata)#系数不显著
fit2.3 <- lm(Y~X1:X2+X2+X3,data=mydata)#系数不显著
fit2.4 <- lm(Y~X1*X2*X3,data=mydata)#系数不显著
```
###model3：
对数线性回归
```
fit3 <- lm(Y~X1+X2+log(X3),data=mydata)#R方0.6512
fit3.1 <- lm(log(Y)~X1+X2+X3,data=mydata)#R方0.559
fit3.2 <- lm(log(Y)~X1+X2+log(X3),data=mydata)#0.5508
```
###model4：
半参数回归（调用不同的包，不同的包由于核函数的选择计算结果可能不同,故只选用了默认的核函数）
将X1,X2(虚拟变量)视为参数部分，X3视为非参数部分进行回归
```
fit4 <- gam(Y~X1+X2+s(X3),data=mydata)
library("np")#这个包没有预测函数?
X <- data.frame(mydata$X1, factor(mydata$X2))
Z <- data.frame(mydata$X3)
bw <- npplregbw(xdat=X, zdat=Z, ydat=mydata$Y)
pl <- npplreg(bws=bw)
```
###model5：
半参数回归2
将X1视为参数部分，X2，X3视为非参数部分进行回归
```
X <- data.frame(mydata$X1)
Z <- data.frame(factor(mydata$X2),mydata$X3)
bw <- npplregbw(xdat=X, zdat=Z, ydat=mydata$Y)
pl <- npplreg(bws=bw)
```