**云南大学数学与统计学院**

**实验报告**

**实验课名称： 应用多元统计分析实验**

**指导教师： 李会琼**

**专业（年级）： 统计学2021级**

**学生姓名： 枫叶 学号:**

**实验名称： 期中测验**

**实验成绩：**

**《应用多元统计分析实验》实验报告 8**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 实验名称 | 期中测验 | | 实验成绩 |  | |
| 学号 |  | | 姓名 | 枫叶 | |
| 实验时间 | 2024年5月14日 | 实验地点 | 格物楼3508 | 指导教师 | 李会琼 |
| 1. **实验目的**   期中测验   1. **实验要求**   1. 对所使用的方法与所得到的结果进行适当的文字描述。  2. 在实验结果的相应部分附上完整的代码与适当的注释。  3. 采用一定的可视化方法体现出对应计算结果。   1. **实验内容**   期中测验   1. **实验软件**   R语言   1. **实验结果**  加载包 library(readxl) library(dplyr) library(purrr) library(CovTools) library(MASS) library(corpcor) source("正态总体假设检验函数汇集.R") source("距离判别函数汇集.R") 第一题 基于如下公式进行估计  给定的情况下的条件均值为  给定的情况下和的条件协方差为  需要指出，上述公式是基于分块矩阵求逆得到的等式，但可以注意到其中条件均值公式和最小二乘回归估计的结果是一致的，问题在于用样本矩替代总体矩的过程，样本协方差阵的逆不是总体协方差阵逆的一个好的估计，虽然在Gauss-Markov条件满足的情况下，这一替代的效果是较好的，但在样本量较小的情况下可能表现不佳甚至无法估计，经查阅资料，发现有基于正态假定的graphic lasso估计方法，该方法实质上是为对数似然添加正则项，此外也有基于稀疏性或带状假定的特殊估计方法，经过尝试，graphic lasso估计结果与直接由样本协方差阵求逆得到的结果有一定差异，但差异并不大。考虑到稳健性与估计难度，下面涉及协方差阵求逆时均使用graphic lasso方法，同时考虑到问题的复杂性，下面只求固定其余所有变量的情况下某一变量的均值和某两个变量的协方差  data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计期中数据.xlsx") %>%  dplyr::select(-1) condition\_mean\_cal <- function(data,x,glasso=T){  #x表示以第几个样本的数据为条件，也可以自行输入一个数据向量  data <- as.matrix(data)  cov\_data <- cov(data)  mean\_data <- apply(data,2,mean)  list1 <- list(type="confidence",param=0.95)  cmean <- c()  for (i in 1:ncol(data)){  data\_i <- data[,(c(i,1:ncol(data))[-i])]  if (length(x)==1){  if (glasso){  cmean <- c(cmean,mean\_data[i]+cov(data\_i)[1,-1]%\*%PreEst.glasso(data\_i[,-1],list1)$C%\*%(data[x,-i]-mean\_data[-i]))  }else{  cmean <- c(cmean,mean\_data[i]+cov(data\_i)[1,-1]%\*%solve(cov(data\_i[,-1]))%\*%(data[x,-i]-mean\_data[-i]))  }  }else{  if (glasso){  cmean <- c(cmean,mean\_data[i]+cov(data\_i)[1,-1]%\*%PreEst.glasso(data\_i[,-1],list1)$C%\*%(x-mean\_data[-i]))   }else{  cmean <- c(cmean,mean\_data[i]+cov(data\_i)[1,-1]%\*%solve(cov(data\_i[,-1]))%\*%(x-mean\_data[-i]))  }  }  }  if (length(x)==1){  cat("当其余变量固定为第",x,"个样本的情况时，条件均值为","\n")  }else{  cat("当其余变量固定为该水平时，条件均值为","\n")  }  return(list("条件均值"=cmean,"样本均值"=as.vector(mean\_data))) } condition\_cov\_cal <- function(data,x){  #x为第几对变量组合，变量组合按1-2,1-3,...,2-3,...顺序排列  data <- as.matrix(data)  cov\_data <- cov(data)  list1 <- list(type="confidence",param=0.95)  cov\_id <- combn(1:ncol(data),2)  cov\_name <- apply(cov\_id,2,paste,collapse="和")  result <- list()  for (i in x){  data\_i <- data[,c(cov\_id[,i],(1:ncol(data))[-cov\_id[,i]])]  cov\_i <- cov(data\_i)  pre\_i <- PreEst.glasso(data\_i[,-c(1,2)],list1)  con\_voc\_i <- cov\_data[1:2,1:2]-matrix(c(cov\_data[1,3:ncol(data)],cov\_data[2,3:ncol(data)]),nrow = 2,byrow = T)%\*%pre\_i$C%\*%matrix(c(cov\_data[3:ncol(data),1],cov\_data[3:ncol(data),2]),ncol = 2)  }  cat("给定其他变量的情况下，变量",cov\_name[x],"的条件协方差为：","\n")  return(list("条件协方差"=con\_voc\_i,"样本协方差"=cov\_data)) } #样例 condition\_mean\_cal(data,3,glasso = T)  ## 当其余变量固定为第 3 个样本的情况时，条件均值为  ## $条件均值 ## [1] 3174.70587 89.54482 1064.12532 564.96063 806.97432 1423.55844 1079.43859 ## [8] -134.57834 ##  ## $样本均值 ## [1] 4385.4064 1319.5400 1246.4118 714.0236 946.1745 1560.9027 1577.9664 ## [8] 497.0136  condition\_cov\_cal(data,2)  ## 给定其他变量的情况下，变量 1和3 的条件协方差为：  ## $条件协方差 ## X1 X2 ## X1 19641.035 -6983.434 ## X2 -6983.434 7701.251 ##  ## $样本协方差 ## X1 X2 X3 X4 X5 X6 X7 ## X1 1744013.51 140322.097 207712.533 284843.79 95496.31 1019893.26 832348.67 ## X2 140322.10 39241.737 8192.824 27732.76 14078.79 101792.68 87622.20 ## X3 207712.53 8192.824 45530.450 29385.07 11469.18 123949.74 87668.19 ## X4 284843.79 27732.760 29385.066 59823.96 25606.08 170226.16 150579.15 ## X5 95496.31 14078.793 11469.179 25606.08 58016.78 42935.43 49738.75 ## X6 1019893.26 101792.678 123949.739 170226.16 42935.43 643156.40 509955.60 ## X7 832348.67 87622.195 87668.189 150579.15 49738.75 509955.60 438022.86 ## X8 237787.46 25434.686 27650.692 42069.44 14376.57 141274.23 114535.16 ## X8 ## X1 237787.46 ## X2 25434.69 ## X3 27650.69 ## X4 42069.44 ## X5 14376.57 ## X6 141274.23 ## X7 114535.16 ## X8 38593.93  事实上，虽然本题样本量较小，但依旧可以直接对协方差阵求逆，直接求逆计算出的结果如下所示  condition\_mean\_cal(data,3,glasso = F)  ## 当其余变量固定为第 3 个样本的情况时，条件均值为  ## $条件均值 ## [1] 3174.70059 89.53364 1064.09182 565.02182 806.99091 1423.60818 1079.37364 ## [8] -134.57273 ##  ## $样本均值 ## [1] 4385.4064 1319.5400 1246.4118 714.0236 946.1745 1560.9027 1577.9664 ## [8] 497.0136  可以看到两种估计方法在协方差阵可逆时差距不大，但矩阵不可逆时直接求逆方法就失效了。开头也已经指出，最小二乘法所求得的也正是某一变量对其他变量的条件均值，并且其形式其实与上面所用的条件均值公式是完全一致的，为简便起见，下面只展示X1的条件均值，可以看到其值与上面所求得的是一致的  predict(lm(X1~.,data))  ## 1 2 3 4 5 6 7 8  ## 5739.080 4712.678 3174.701 3134.059 3410.811 4362.869 3544.398 3055.831  ## 9 10 11  ## 7168.126 4534.183 5402.734 第二题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计期中数据.xlsx",sheet = 2) %>%  dplyr::select(-c(1,2)) newdata <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计期中数据.xlsx",sheet = 3) %>%  dplyr::select(-c(1,2)) multi.cov.test(data[,-9],data$group,3)  ## $p.value ## [1] 0 ##  ## $correct.M ## [1] 318.9236  distinguish.distance.correct(data[,-9],factor(data$group),newdata)  ## Estimating optimal shrinkage intensity lambda.var (variance vector): 0.4867  ##  ## Estimating optimal shrinkage intensity lambda (correlation matrix): 0.5108  ## 1 2 ## blong 1 1  经检验，认为各总体协方差阵不同，但进一步计算发现存在协方差阵奇异问题，使用shrink估计修正过的判别函数作判别，结果显示新样本 第三题 data <- read\_xlsx("D:/预删除文件夹/大三下/多元统计/多元统计期中数据.xlsx",sheet = 4) %>%  dplyr::select(-1) %>%  mutate(across(6,factor,labels=c(1,2)))  ## Warning: There was 1 warning in `mutate()`. ## ℹ In argument: `across(6, factor, labels = c(1, 2))`. ## Caused by warning: ## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0. ## Supply arguments directly to `.fns` through an anonymous function instead. ##  ## # Previously ## across(a:b, mean, na.rm = TRUE) ##  ## # Now ## across(a:b, \(x) mean(x, na.rm = TRUE))  multi.cov.test(data[,-6],data$学校,2)  ## $p.value ## [1] 0.9794409 ##  ## $correct.M ## [1] 3.081483  data1 <- filter(data,学校==1) %>% dplyr::select(-6) data2 <- filter(data,学校==2) %>% dplyr::select(-6) two.mu.test(data1,data2)  ## $统计量 ## [1] 1.508911 ##  ## $临界值 ## [1] 2.958249 ##  ## $p.value ## [1] 0.2493143  经检验，可以认为两总体协方差阵相同，基于此假设进一步作两总体均值检验，结果显示不能拒绝均值相等的原假设 | | | | | |