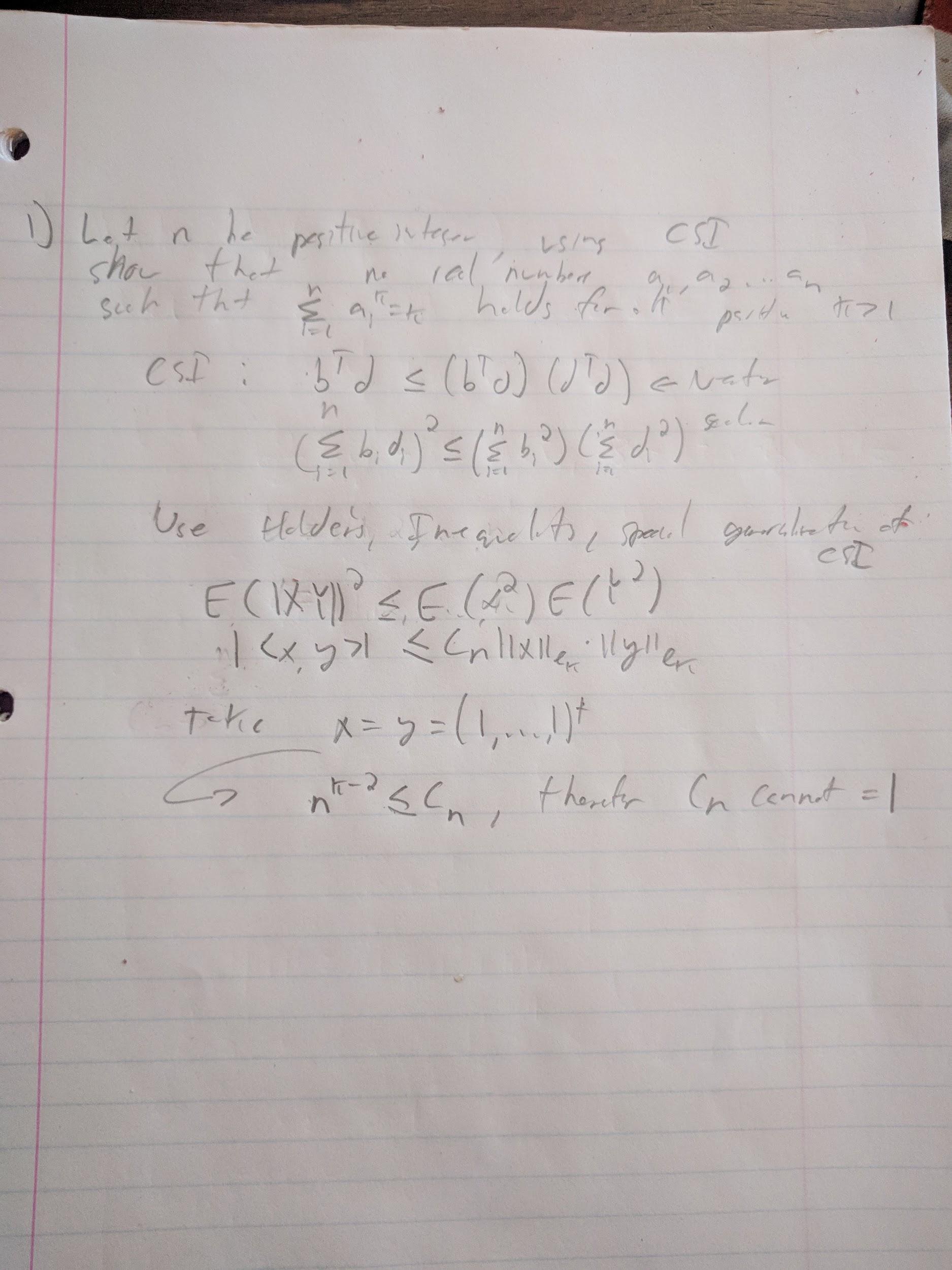
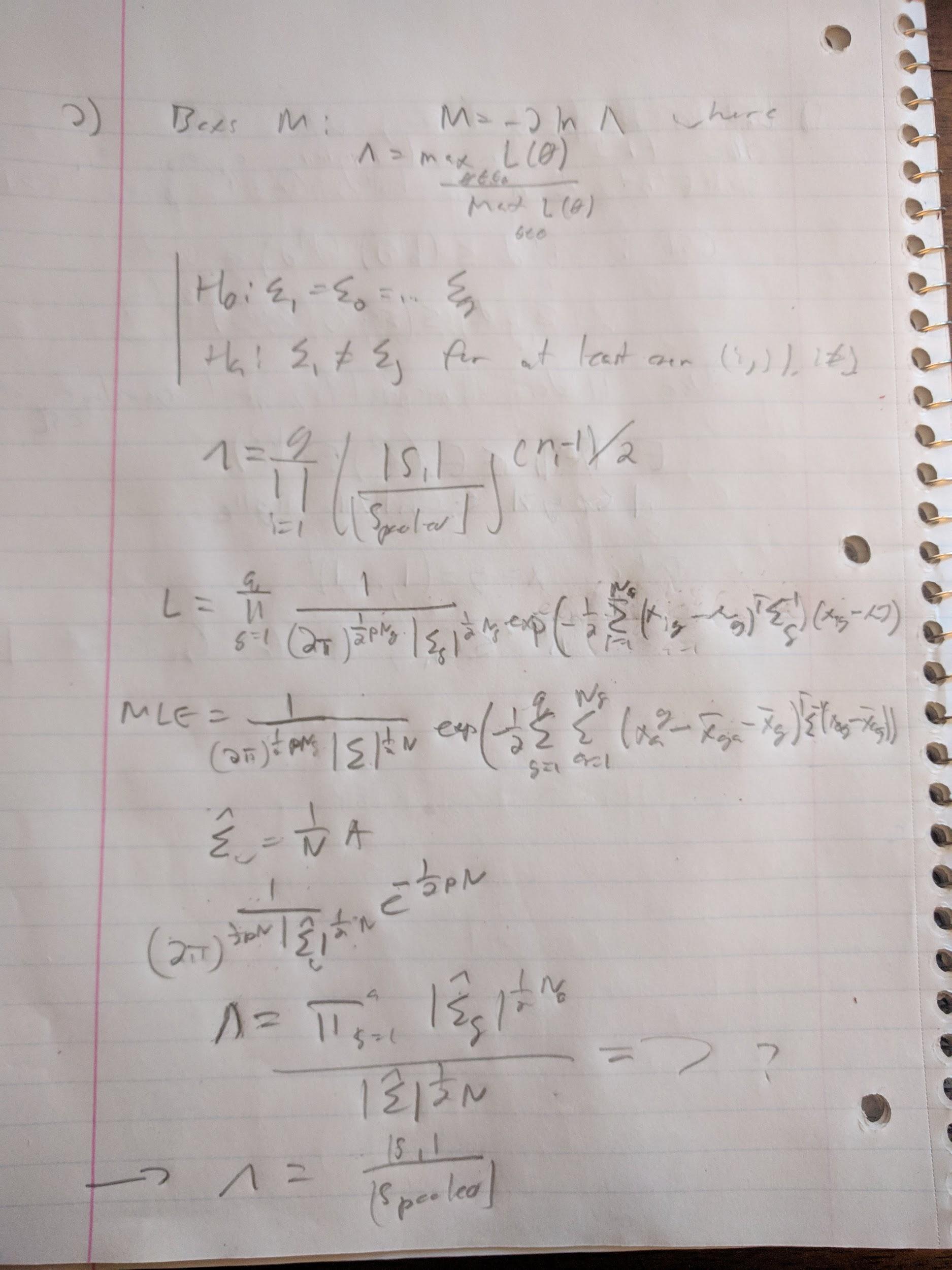
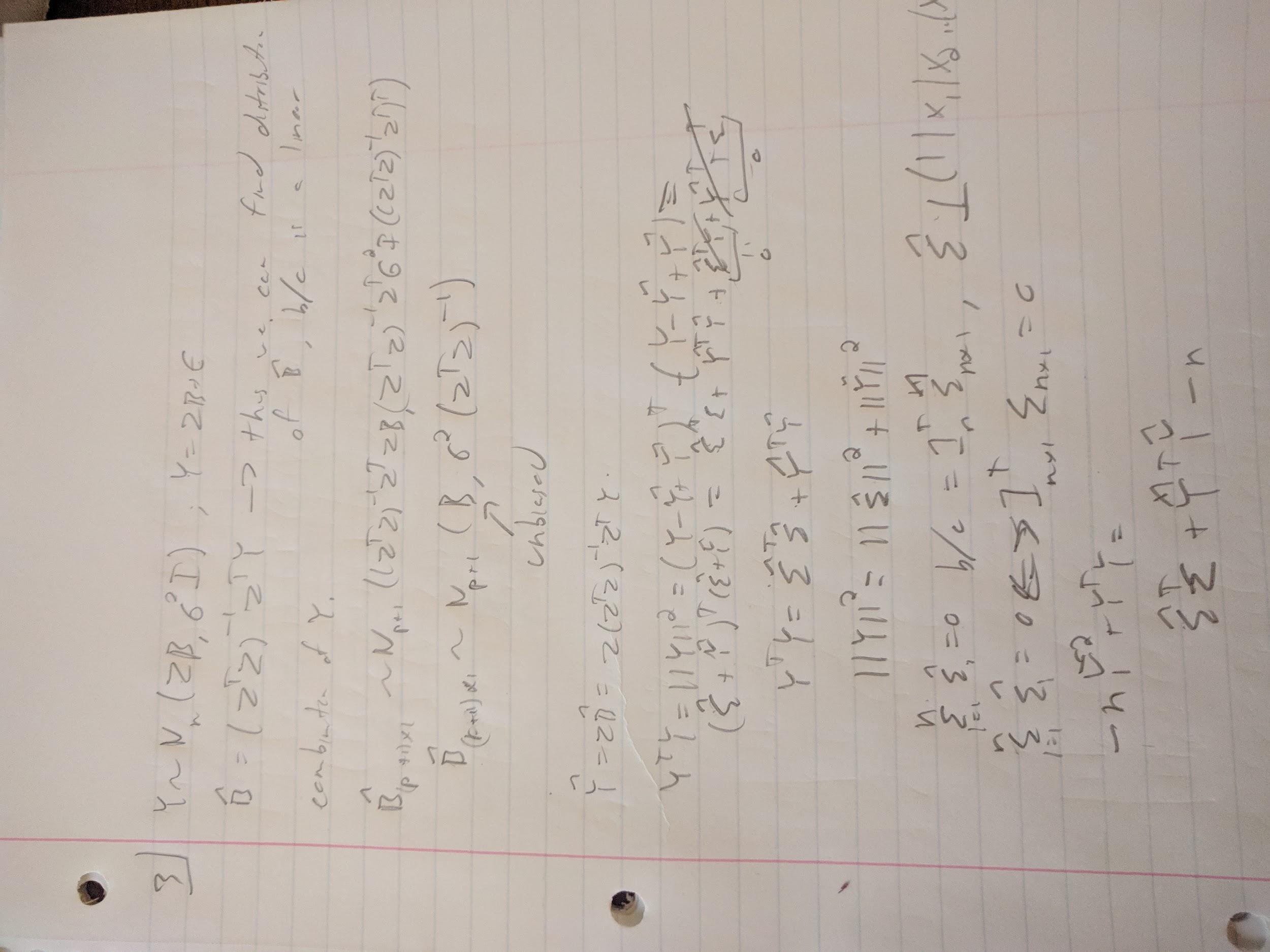
**Problem 1**

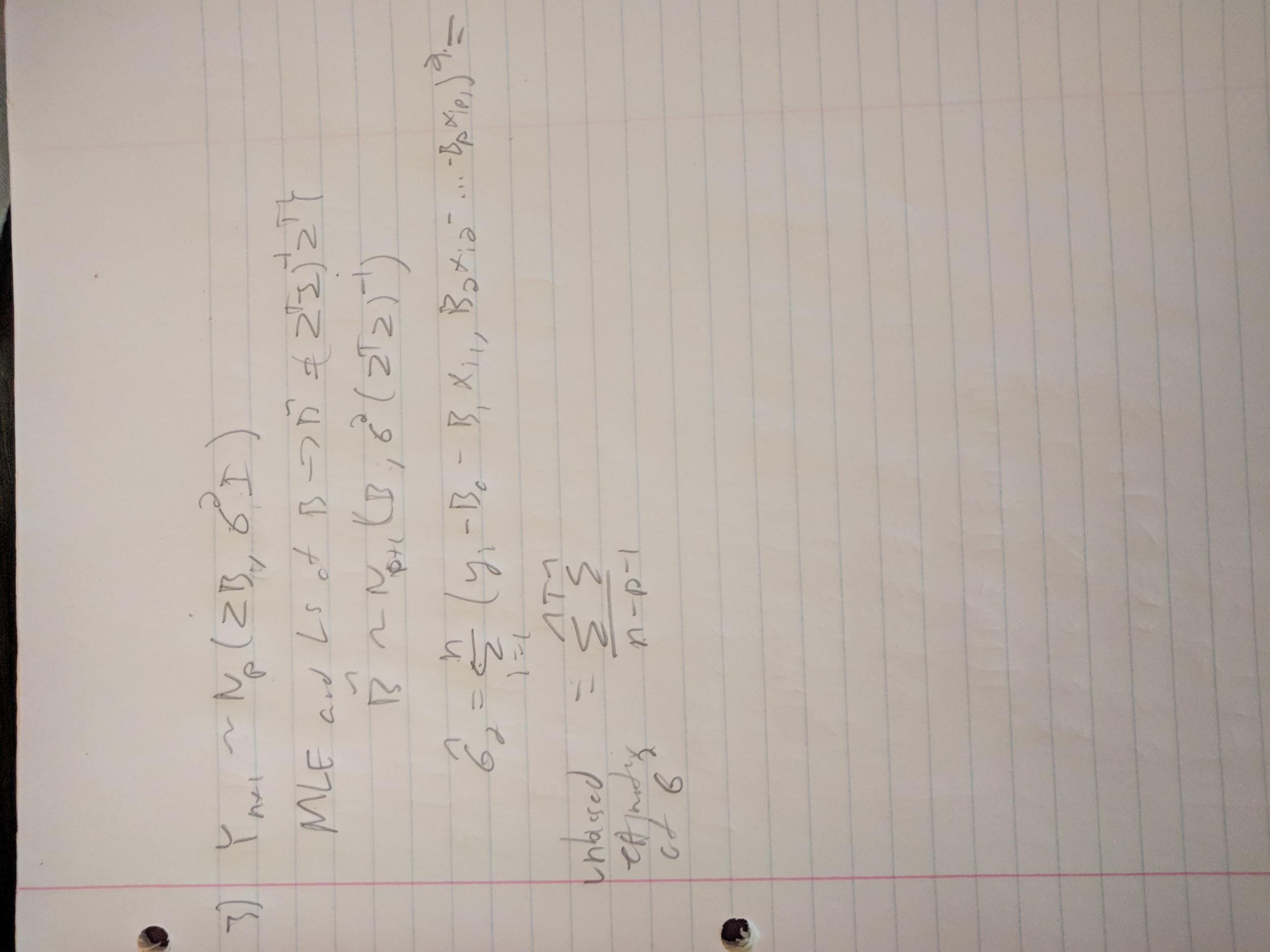
****

**Problem 2**

****

**Problem 3**

****

****

**Problem 4**

**Code:**

library(MASS)

library(car)

# set output file

sink("pension\_report.txt")

# read data

pension = as.data.frame(read.table("Pension.txt", head=TRUE))

# read first 6 rows

head(pension)

# apply(pension, 2, is.na)

# remove na

# pension = na.omit(pension)

# get dimensions

dim(pension)

# read names

names(pension)

# get summary stats

summary(pension)

pension.clean <- na.omit(pension)

# histogram and boxplot each col

for(col in 1:ncol(pension)){

print(names(pension)[col])

png(file=paste("hist\_",names(pension)[col],".png"))

hist(pension[,col], main=names(pension)[col], xlab=names(pension)[col])

dev.off()

# skip box plotting wealth

if((names(pension)[col]) != "wealth89"){

png(file=paste("boxplot\_",names(pension)[col],".png"))

f = as.formula(paste("wealth89 ~ ",names(pension)[col]))

boxplot(f, data=pension, main=names(pension)[col], xlab=names(pension)[col])

dev.off()

}

}

#png(file="scatterplot\_small.png")

# some options to create a higher res scatterplot

#png(file="scatterplot", height=10, width=10, units="in", res=1000, pointsize=4)

#par(

# mar = c(5, 5, 2, 2),

# xaxs = "i",

# yaxs = "i",

# cex.axis = 2,

# cex.lab = 2

#)

scatterplotMatrix(pension)

#dev.off()

base <- lm(formula=wealth89 ~ .,data = pension.clean)

step <- stepAIC(base, trace = FALSE)

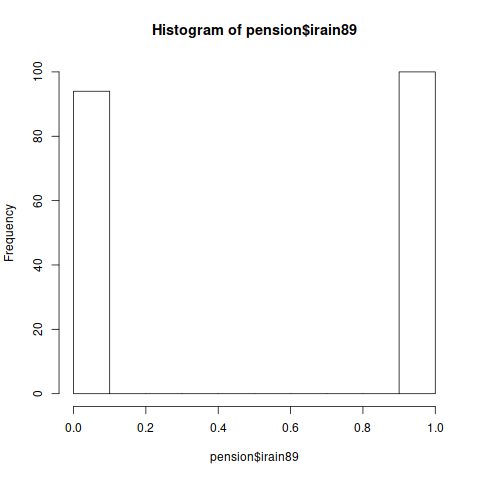
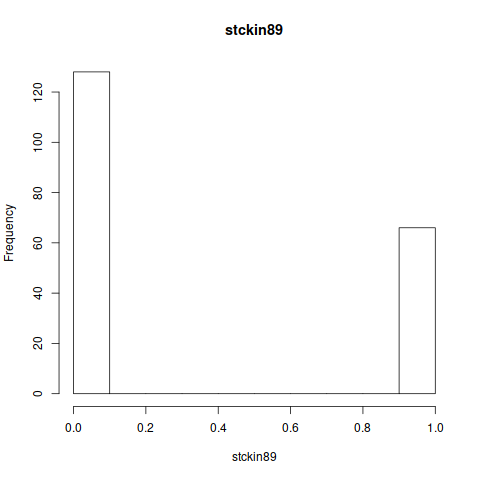
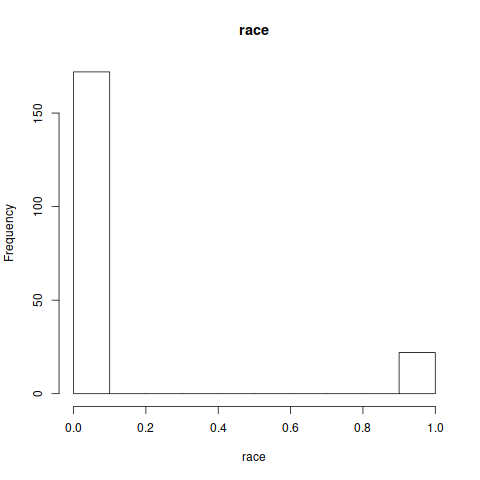
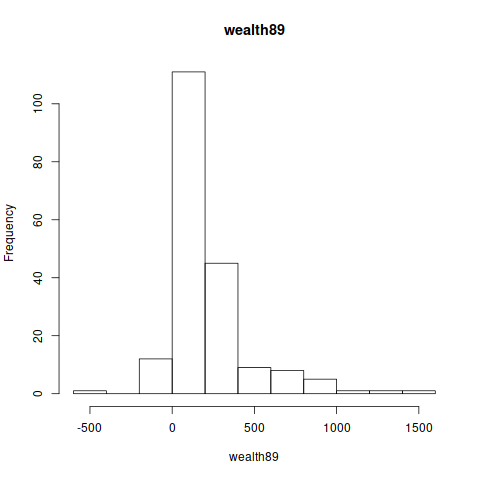
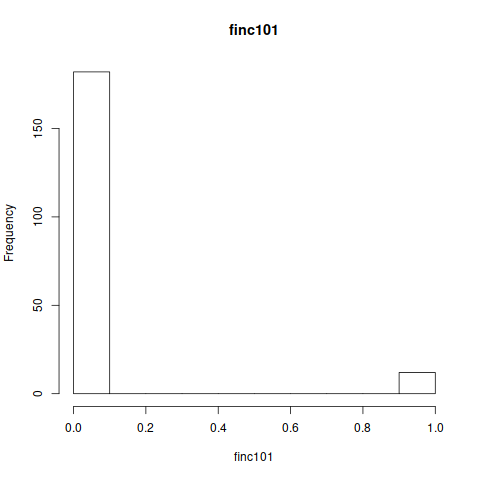
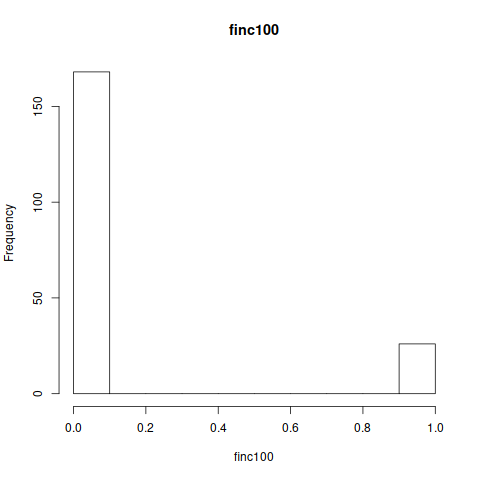
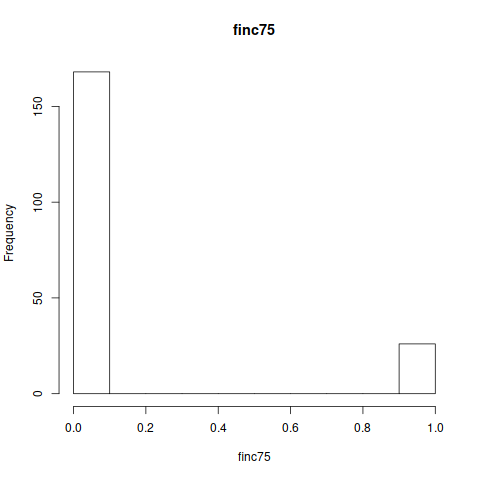
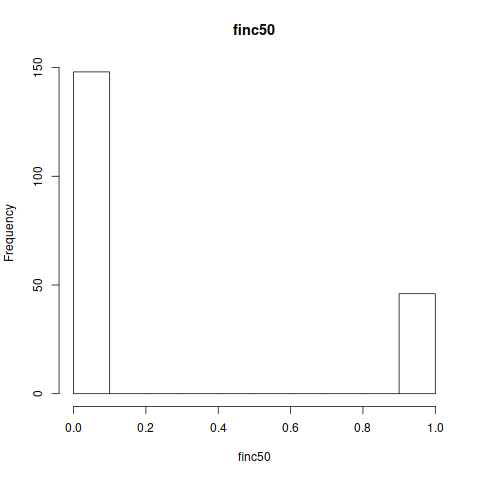
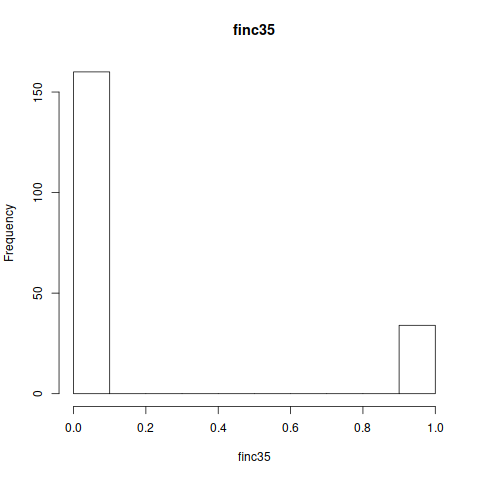
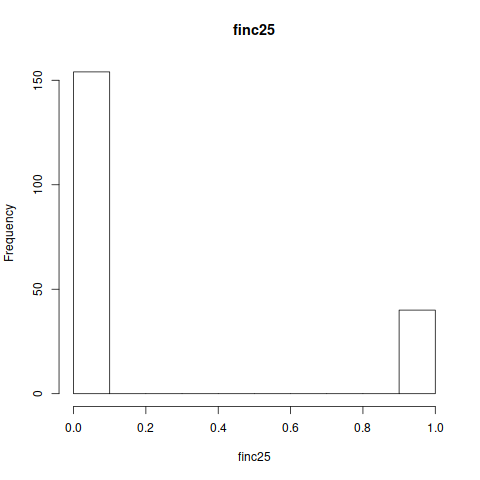
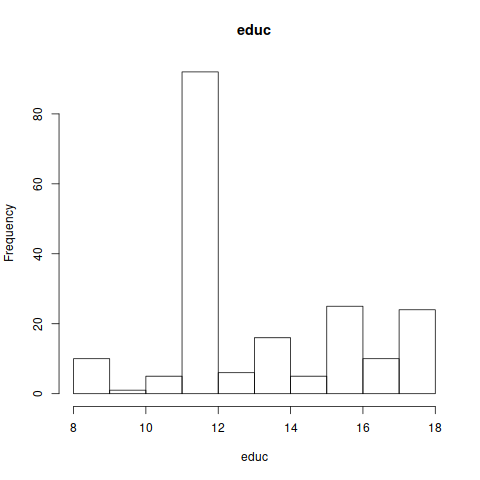
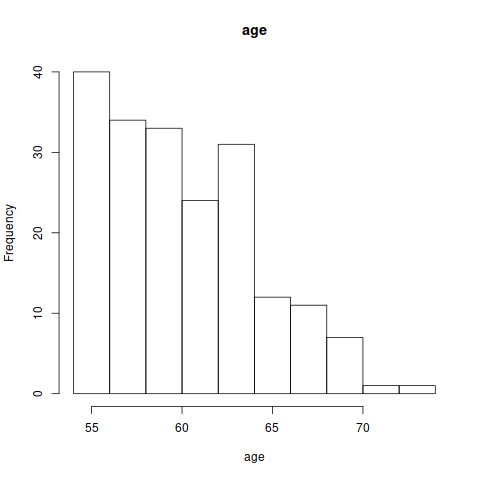
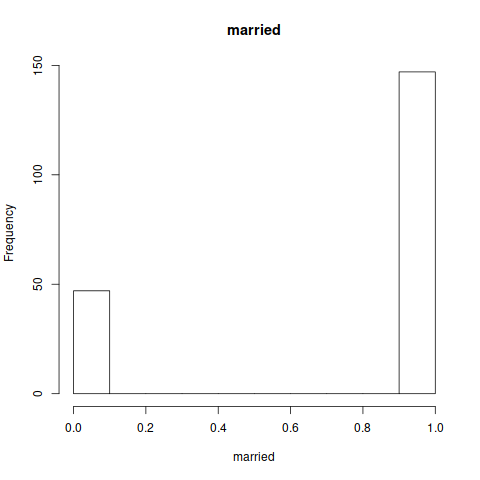
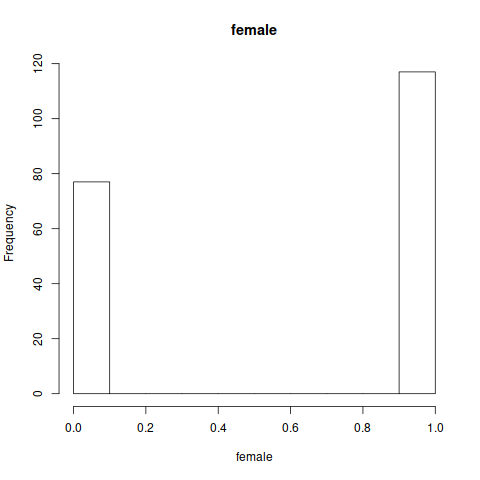
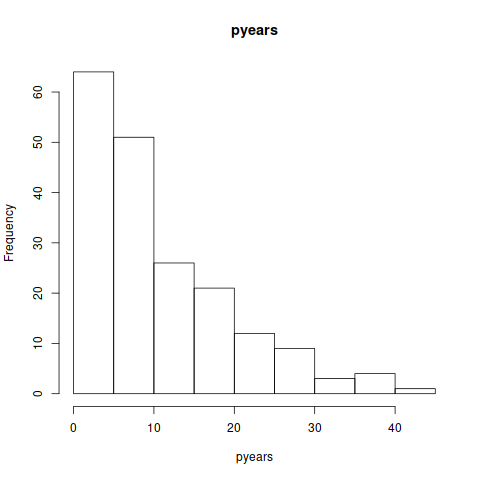
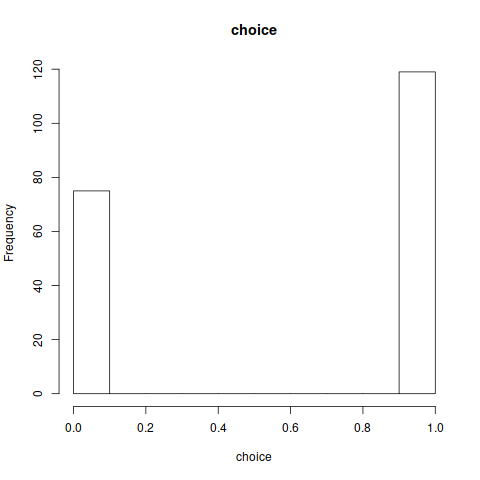
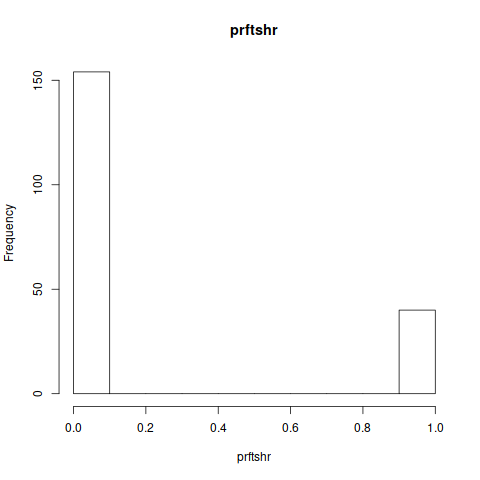
step$anova

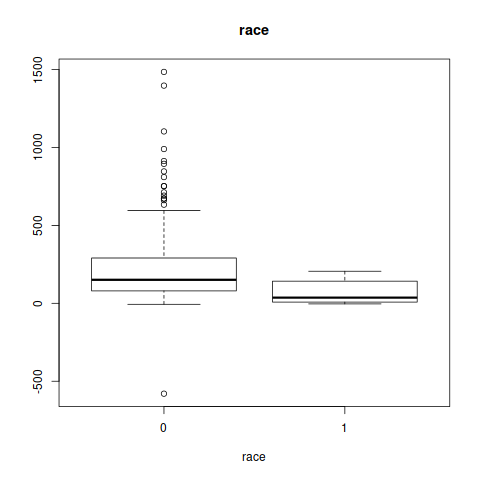
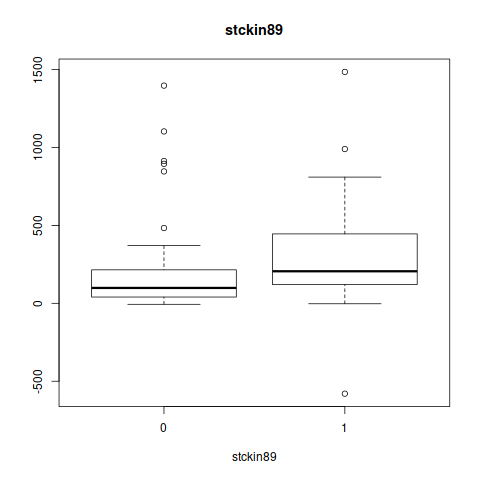
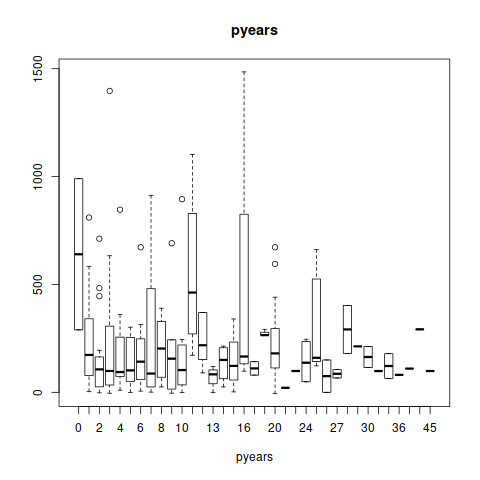
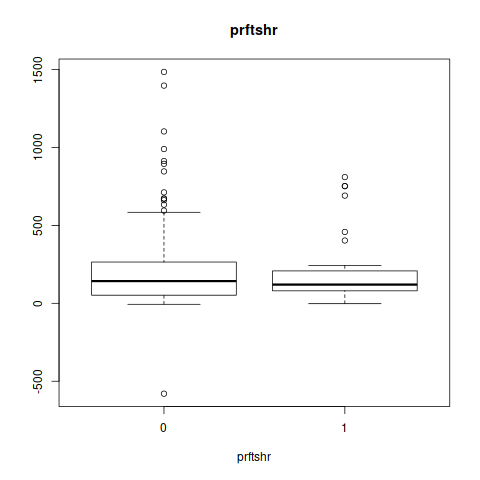
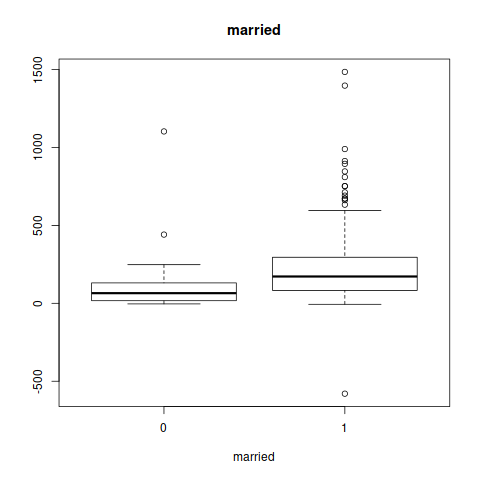
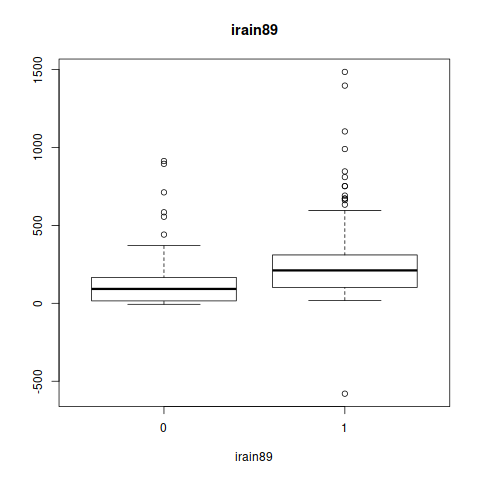
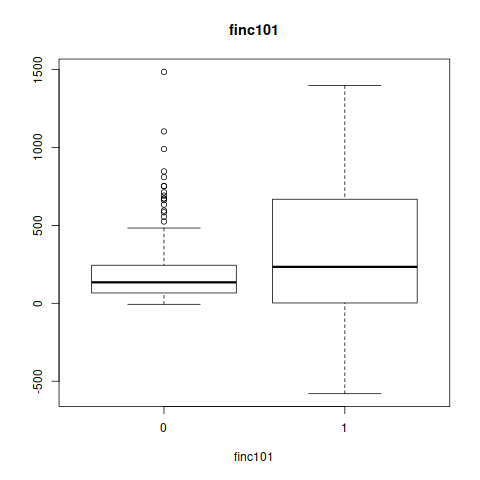
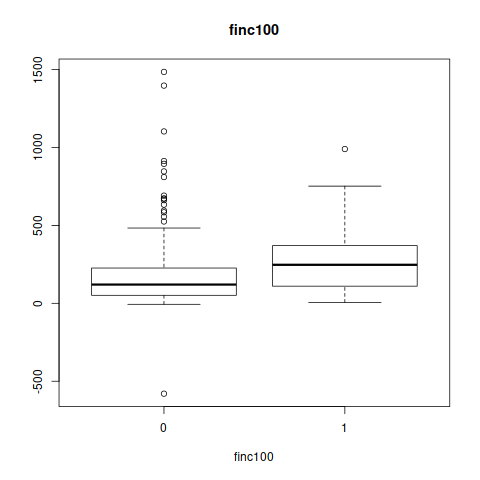
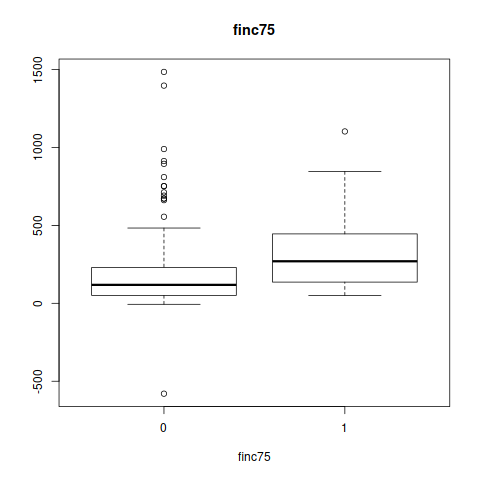
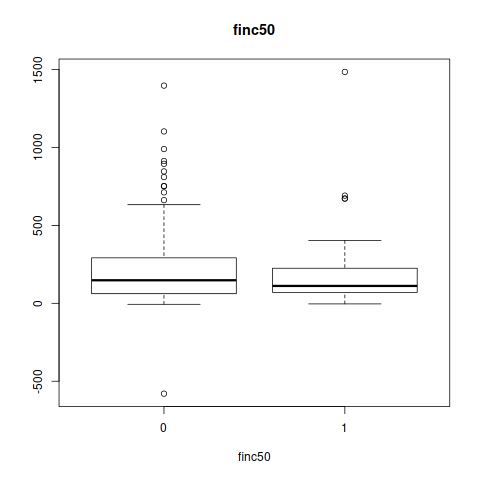
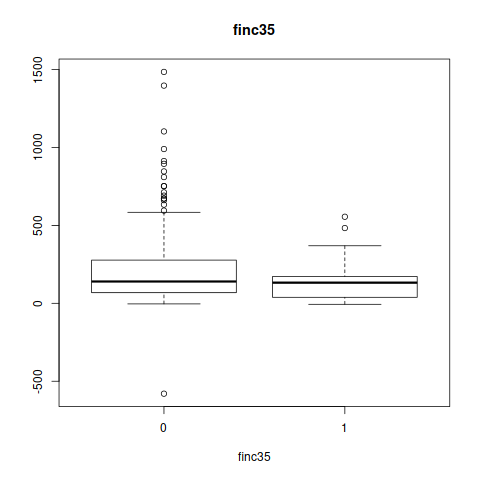
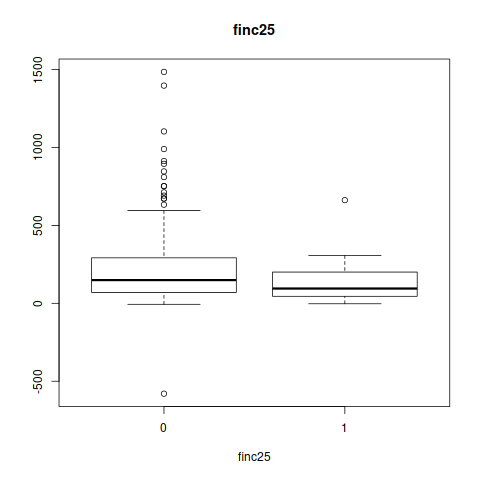
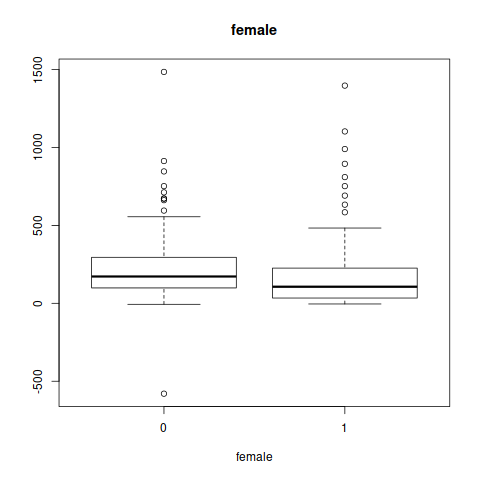
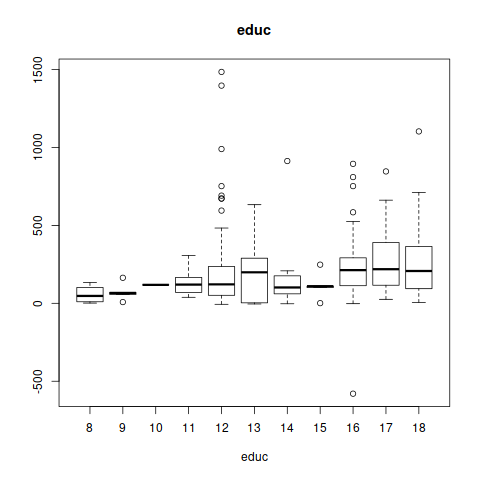
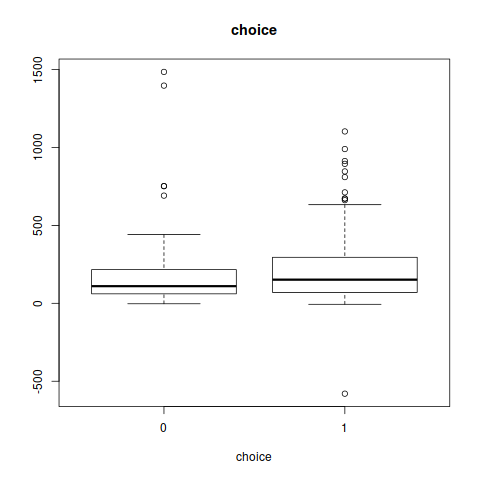
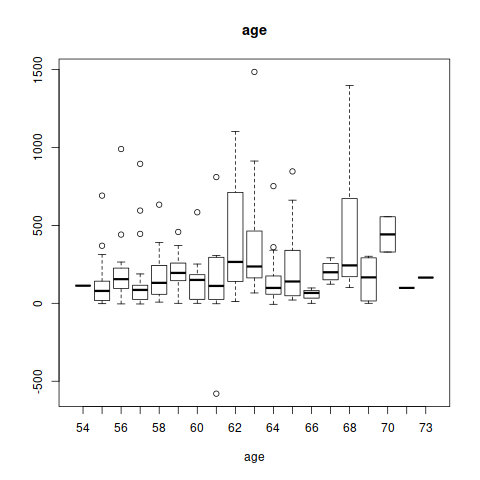
step2 <- stepAIC(base, ~ .^2 + I(age)^2, trace = FALSE)

step2$anova

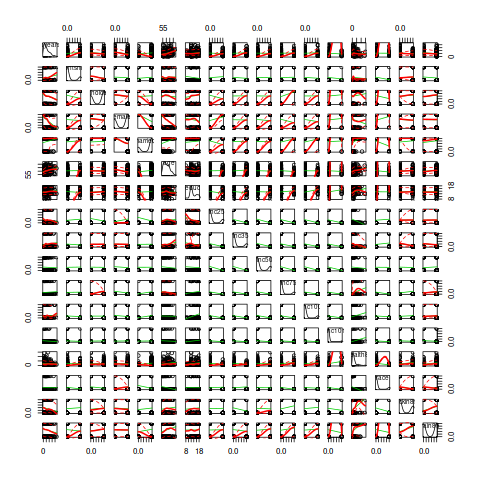
# unset sink file

sink()





From the scatterplot matrix produced, it would appear that the factor most highly correlated with wealth89 is the irain89 factor, which is expected as IRA allows for wealth to accumulate tax free. Stockin89 is also a positive correlation to wealth. Finc101 is also positively correlates. with the other finc factors also being positively correlated, but to a much lesser extent. Notably, there is little correlation between pyears and wealth. Married is positively correlated with all of the finc metrics as well. Additionally, female is negatively correlated with pyears as might be expected as many mothers do not work. Another notable item is that race is positively correlated with wealth. Pyears correlates positively to profit sharing and company contribution choices.



**Base Model Fitting**

Fitting a linear model to the data using all variables produces a model with:

Multiple R-squared: 0.317, Adjusted R-squared: 0.2542. We can improve this by using stepwise AIC to find better combinations of variables.

**AIC Minimization**

Initial Model:

wealth89 ~ pyears + prftshr + choice + female + married + age +

educ + finc25 + finc35 + finc50 + finc75 + finc100 + finc101 +

race + stckin89 + irain89

Final Model:

wealth89 ~ age + finc50 + finc75 + finc100 + finc101 + stckin89 +

irain89

Step Df Deviance Resid. Df Resid. Dev AIC

1 174 7887254 2064.041

2 - finc25 1 603.2244 175 7887857 2062.055

3 - prftshr 1 7635.5614 176 7895493 2060.240

4 - female 1 17065.4348 177 7912558 2058.652

5 - finc35 1 16678.4486 178 7929237 2057.055

6 - race 1 24227.8331 179 7953464 2055.637

7 - choice 1 25780.8685 180 7979245 2054.255

8 - pyears 1 52375.0676 181 8031620 2053.505

9 - educ 1 52353.1422 182 8083973 2052.746

10 - married 1 70340.5608 183 8154314 2052.401

Notably, this is a much more compact model, and the factors in the final model are most of the same that are noted in the preceding scatterplot analysis revealed. However, the AIC only decreases by a very small amount, less than 1%. Additionally, the residual deviance actually increases from the first step to the tenth, increasing from 7887254 to 8154314. R squared for this updated model is: Multiple R-squared: 0.2938, Adjusted R-squared: 0.2668, neither of which are very good, the adjusted is slight better than using all variables and the multiple worse.

**Stepwise AIC with combinations and age with itself**

Initial Model:

wealth89 ~ pyears + prftshr + choice + female + married + age +

educ + finc25 + finc35 + finc50 + finc75 + finc100 + finc101 +

race + stckin89 + irain89

Final Model:

wealth89 ~ pyears + prftshr + choice + female + married + age +

educ + finc50 + finc75 + finc100 + finc101 + race + stckin89 +

irain89 + married:finc75 + age:finc101 + female:finc101 +

pyears:finc50 + finc101:irain89 + female:irain89 + prftshr:finc100 +

female:finc100 + female:age + pyears:age + choice:finc50 +

prftshr:finc75 + prftshr:finc101 + educ:finc101 + pyears:finc101 +

choice:finc101 + finc101:stckin89 + married:stckin89 + age:finc100 +

educ:finc50 + race:irain89

Step Df Deviance Resid. Df Resid. Dev AIC

1 174 7887254 2064.041

2 + married:finc75 1 682092.8826 173 7205161 2048.765

3 + age:finc101 1 680633.2340 172 6524528 2031.812

4 + female:finc101 1 217110.5875 171 6307417 2027.348

5 + pyears:finc50 1 192372.7008 170 6115044 2023.432

6 + finc101:irain89 1 165214.5480 169 5949830 2020.201

7 + female:irain89 1 177354.1872 168 5772476 2016.421

8 + prftshr:finc100 1 169299.4394 167 5603176 2012.735

9 + female:finc100 1 124103.8779 166 5479072 2010.457

10 + finc25:stckin89 1 119879.8634 165 5359193 2008.232

11 - finc35 1 11579.3336 166 5370772 2006.644

12 + female:age 1 90718.5811 165 5280053 2005.390

13 + pyears:age 1 107506.1432 164 5172547 2003.461

14 + choice:finc50 1 100505.5096 163 5072042 2001.713

15 + married:race 1 93352.9869 162 4978689 2000.165

16 + prftshr:finc75 1 81023.7857 161 4897665 1999.031

17 + prftshr:finc101 1 67101.8420 160 4830563 1998.396

18 + educ:finc101 1 119761.2339 159 4710802 1995.601

19 + pyears:finc101 1 115194.1986 158 4595608 1992.873

20 + choice:finc101 1 250213.9384 157 4345394 1984.180

21 + finc101:stckin89 1 187021.0505 156 4158373 1977.777

22 + married:stckin89 1 57052.1233 155 4101321 1977.138

23 - finc25:stckin89 1 22043.2767 156 4123364 1976.162

24 - finc25 1 631.1075 157 4123995 1974.191

25 + age:finc100 1 64938.6330 156 4059056 1973.160

26 + educ:finc50 1 66615.7725 155 3992440 1971.999

27 - married:race 1 39158.8547 156 4031599 1971.863

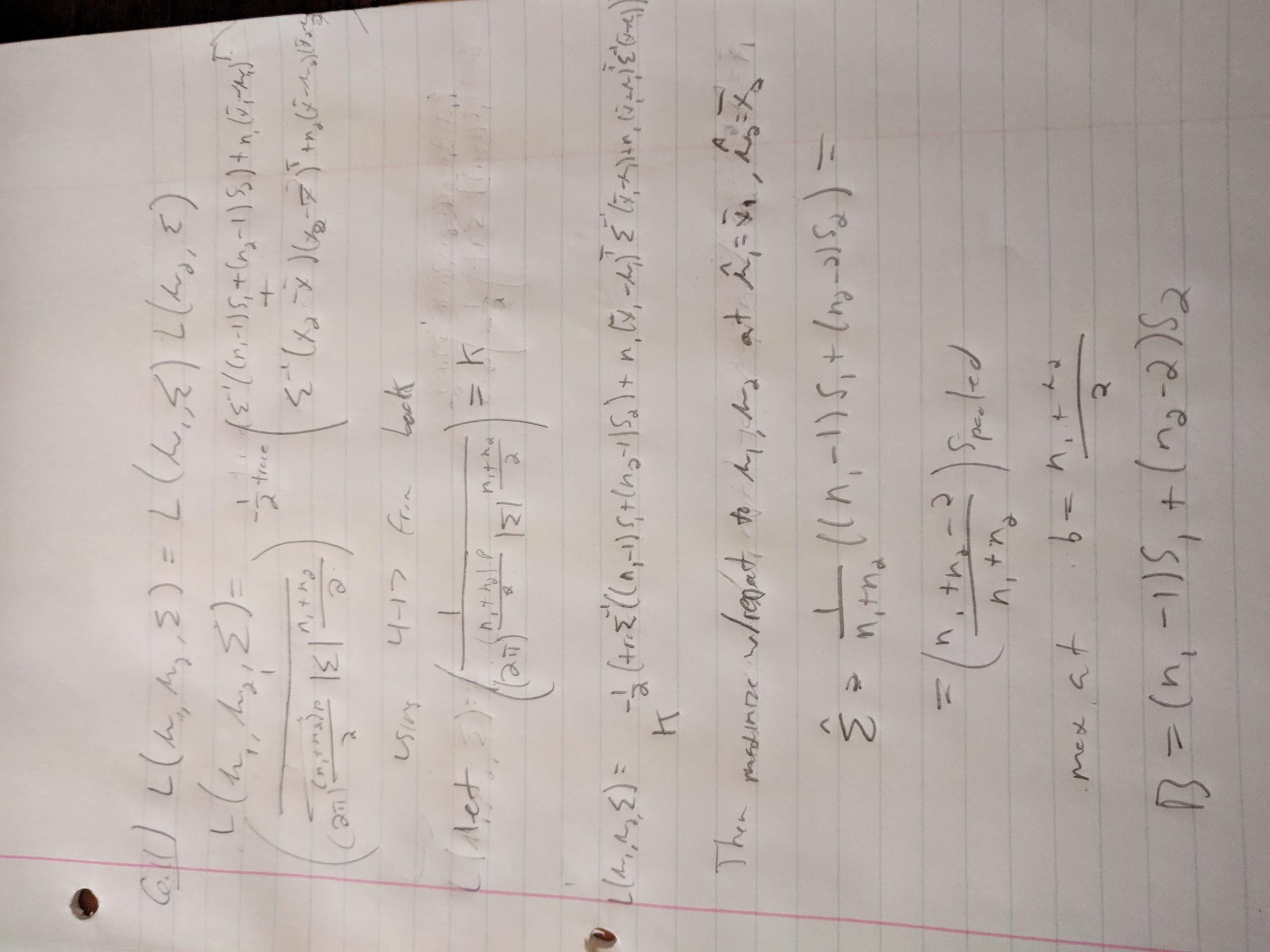
28 + race:irain89 1 44542.2414 155 3987057 1971.741

This stepwise AIC model with combinations and age squared provides a much better model, giving a 4% increase in AIC. Here we can see that the residual deviance decreases from 7887254 to 3987057 - a decrease of nearly 50%. Fitting this refined model yields a multiple R^2 of 0.6547 and an adjusted R^2 of 0.5768, an improvement over the base and initial AIC fit by a very large margin, this is a good model.

Looking at the P-Values from both refined models, we can see that the finc101 appears in quite a high number of the combinatorial factors. Other factors occurring frequently are prftshr, choice, irain89, finc100 and finc75. Interestingly, though these are a similar set to those initially identified, they only truly manifest themselves as meaningful when used in conjunction with other predictor variables. It would also appear that finc75 is the most impactful of the finc variables on it’s own - quite possibly as it can be seen as a proxy variable for both finc100 and 101. It seems that the most important factors in predicting wealth at retirement are indeed the ability of an individual to levels of retirement contribution, specifically those at 75, 100, and 101. Additionally, the ability to profit share and to contribute to an IRA are also key factors.

It would be of use to add variables such as salary levels to indicate income streams as well. Another useful variable would be number of dependents, as this could have an effect on the ability to save for retirement. Other useful variables could be cost of living in the area of the individual, or medical expenses. Some feature engineering techniques that could be of use could be PCA, with specific respect to the finc variables, to find the composite of all the levels specifically. We have already applied some feature engineering techniques to remove NA values, and could also use dummy variable for the categorical features in the data.

**Problem 5**



**Problem 6**

d = read.table("T6-8.dat")

#header = c("WordDiff", "WordSame", "ArabicDiff", "ArabicSame")

#names(d) <- header

xbar = colMeans(d)

S=var(d)

C = cbind(rep(1,3), diag(rep(-1,3)))

q = dim(d)[2]

n = dim(d)[1]

T2 = t(C%\*%xbar)%\*%solve(C%\*%S%\*%t(C)/n)%\*%(C%\*%xbar)

# alpha = 0.05

mes = (n-1)\*(q-1)/(n-q+1)\*qf(0.95,q-1,n-q+1)

# reject H0

if (T2>mes) cat("Reject Ho \n") else ("Fail to reject Ho \n")

# univariate T tests for treatment effects

#but what is cause? Well, lets do univariate T tests to find out

t.test(d$V1,d$V2, paired=T)

t.test(d$V1,d$V3, paired=T)

t.test(d$V1,d$V4, paired=T)

t.test(d$V2,d$V3, paired=T)

t.test(d$V2,d$V4, paired=T)

t.test(d$V3,d$V4, paired=T)

contrast.of.subject <- function(x){

c1 = (x[3]+x[4]) - (x[1] + x[2])

c2 = (x[1]+x[3]) - (x[2] + x[4])

c3 = (x[1]+x[4]) - (x[2] + x[3])

return(c(c1,c2,c3))

}

contrast.m <- apply(d,1,contrast.of.subject)

contrast.tm = t(contrast.m)

contrast.tm

Reject Ho

Paired t-test

data: d$V1 and d$V2

t = 6.2609, df = 31, p-value = 5.854e-07

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

61.9988 121.9075

sample estimates:

mean of the differences

91.95312

Paired t-test

data: d$V1 and d$V3

t = 6.0363, df = 31, p-value = 1.107e-06

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

94.18733 190.31267

sample estimates:

mean of the differences

142.25

Paired t-test

data: d$V1 and d$V4

t = 11.018, df = 31, p-value = 3.015e-12

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

209.1197 304.1303

sample estimates:

mean of the differences

256.625

Paired t-test   
   
data: d$V2 and d$V3   
t = 2.5385, df = 31, p-value = 0.01637   
alternative hypothesis: true difference in means is not equal to 0   
95 percent confidence interval:   
 9.88743 90.70632   
sample estimates:   
mean of the differences   
 50.29688   
   
   
 Paired t-test   
   
data: d$V2 and d$V4   
t = 10.853, df = 31, p-value = 4.382e-12   
alternative hypothesis: true difference in means is not equal to 0   
95 percent confidence interval:   
 133.7258 195.6179   
sample estimates:   
mean of the differences   
 164.6719   
   
   
 Paired t-test   
   
data: d$V3 and d$V4   
t = 7.2841, df = 31, p-value = 3.392e-08   
alternative hypothesis: true difference in means is not equal to 0   
95 percent confidence interval:   
 82.35042 146.39958   
sample estimates:   
mean of the differences   
 114.375   
   
 V3 V1 V1   
 [1,] -437.5 98.5 -81.5   
 [2,] -533.0 139.0 101.0   
 [3,] -327.5 132.5 118.5   
 [4,] -464.0 332.0 12.0   
 [5,] -249.0 161.0 109.0   
 [6,] 75.0 331.0 -194.0   
 [7,] -377.0 436.0 116.0   
 [8,] -879.5 -34.5 229.5   
 [9,] -265.0 321.0 -39.0   
[10,] -456.0 252.0 44.0   
[11,] -154.0 132.0 -28.0   
[12,] -405.0 279.0 -103.0   
[13,] -461.5 177.5 -85.5   
[14,] -106.5 107.5 30.5   
[15,] -299.0 182.0 24.0   
[16,] -441.5 505.5 133.5   
[17,] -232.5 149.5 -77.5   
[18,] -35.0 84.0 -95.0   
[19,] -205.0 30.0 -36.0   
[20,] -389.0 227.0 -55.0   
[21,] -259.0 137.0 -123.0   
[22,] -226.0 98.0 -20.0   
[23,] -270.5 211.5 -191.5   
[24,] -12.0 20.0 -88.0   
[25,] -141.5 114.5 -55.5   
[26,] -228.0 149.0 -23.0   
[27,] -740.5 269.5 29.5   
[28,] -46.5 180.5 -5.5   
[29,] -376.5 296.5 -46.5   
[30,] -283.0 353.0 -217.0   
[31,] -147.5 123.5 -45.5   
[32,] -448.0 607.0 -55.0

We can say that yes a multivariate normal distribution is a reasonable model for this data because the resulting data comes from a linear transformation of another multivariate normal distributed data.

**Problem 7**

df = read.table("T8-5.DAT")

df.pca.cor = prcomp(df, scale=TRUE, center=TRUE, retx=TRUE)

df.pca.cov = prcomp(df, scale=FALSE, center=TRUE, retx=TRUE)

png(file="dfpcacor.png")

plot(df.pca.cor, type="l")

dev.off()

summary(df.pca.cor)

png(file="biplot\_cor.png")

biplot(df.pca.cor)

dev.off()

png(file="dfpcacov.png")

plot(df.pca.cov, type="l")

dev.off()

summary(df.pca.cov)

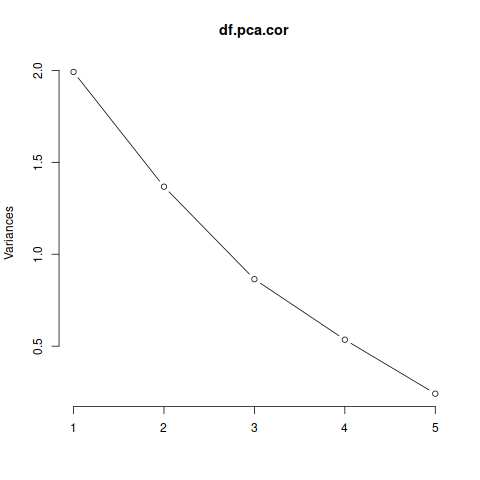
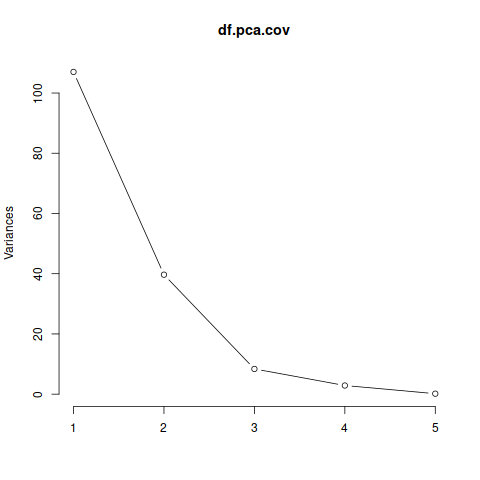
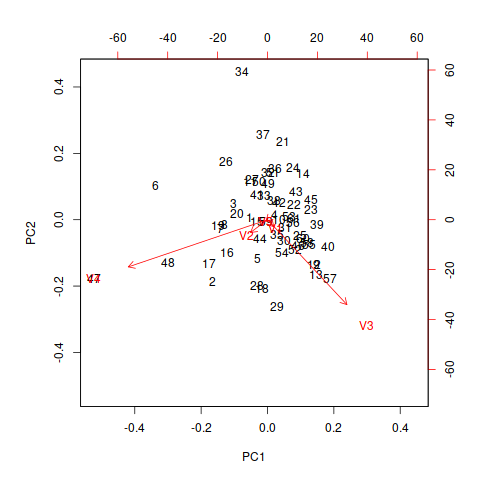
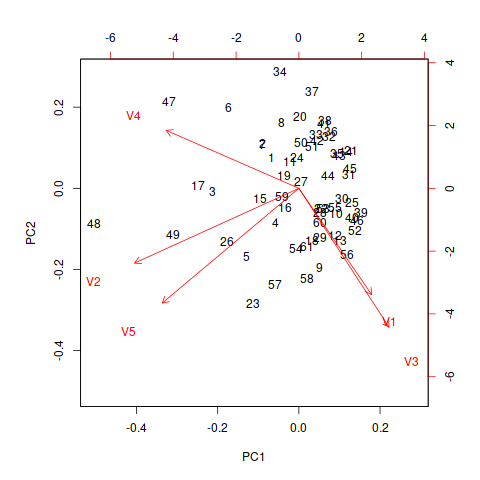
png(file="biplot\_cov.png")

biplot(df.pca.cov)

dev.off()

We do note that we can achieve our objective with both methods, however we can use less PCs with covariance as it does not scale our data. Thus, it can be summarized in as few as 3 dimensions with the covariance method. The principal components have been plotted for interpretation - the magnitude of PCs 1 and 2 is most notable on the covariance method.

Correlation BiPlot Covariance BiPlot PCA Covariance PCA Correlation

****