Natan Alper 3/18/2020

Business Intelligence & Consumer Insights- Professor Kovtun

HW #5

**1 (a)**

x1 = rexp(4000,10)

x2 = rgamma(4000,10)

x3 = 4 - 2\*x2+rnorm(4000,sd=4)

y = 1000 - 2\*x1 - 5\*x2^(-.2) -4\*x3 + rnorm(4000,sd=10)

data = data.frame(x1,x2,x3,y)

par(mfrow=c(3,4))

plot(x2^1.9,y)

plot(x2^(1.5),y)

plot(x2^(1.2),y)

plot(x2^(.8),y)

plot(x2^(.6),y)

plot(x2^(.2),y)

plot(x2^(-.2),y)

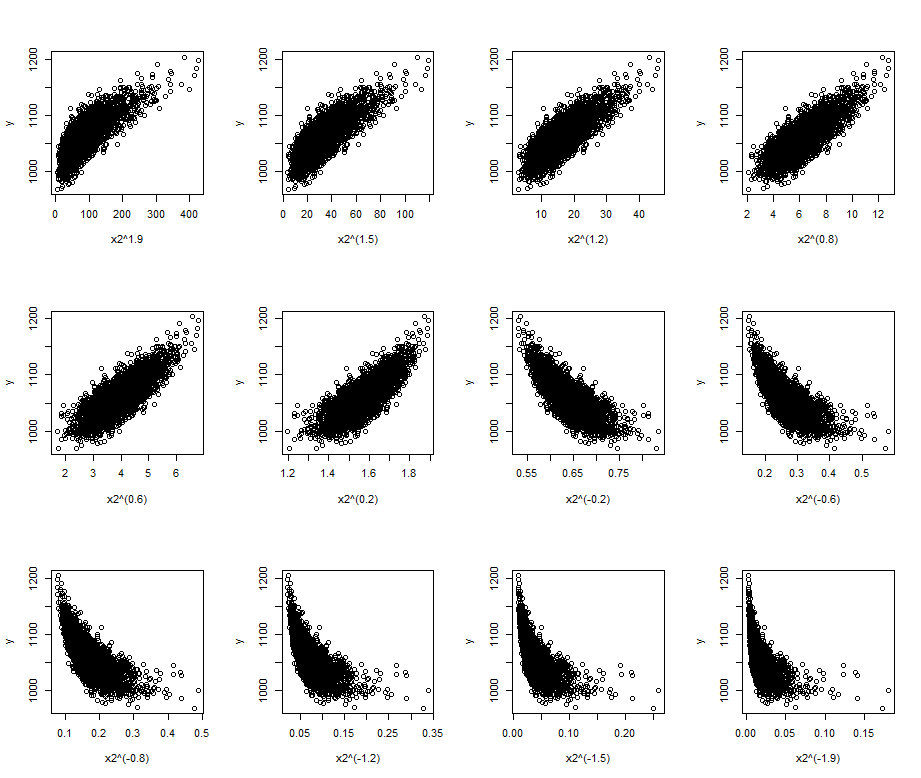
plot(x2^(-.6),y)

plot(x2^(-.8),y)

plot(x2^(-1.2),y)

plot(x2^(-1.5),y)

plot(x2^(-1.9),y)



X2^.8 looks best.

**1 (b)**

library(MASS)

boxcox(lm(y~x2), plotit = FALSE) # Find largest y-value. The corresponding x-value is lamda

# Largest value is where x = 2

lam <- 2

tx2 <- ((x2^lam)-1)/2

lm(y ~ x1 + tx2 + x3)

Call:

lm(formula = y ~ x1 + tx2 + x3)

Coefficients:

(Intercept) x1 tx2 x3

996.367351 -2.588933 -0.007253 -4.051262

Regression Model:

Y = 996.367351 + (-2.588933\*x1) + (-0.007253\*tx2) + (-4.051262\*x3)

**1 (c)**

x1cent <- x1-mean(x1)

x2cent <- x2-mean(x2)

x3cent <- x3-mean(x3)

cor(data.frame(x1cent, x2cent, x3cent))

x1cent x2cent x3cent

x1cent 1.00000000 0.01219508 -0.01337701

x2cent 0.01219508 1.00000000 -0.85251102

x3cent -0.01337701 -0.85251102 1.00000000

pca <- prcomp(data)

summary(pca)

Importance of components:

PC1 PC2 PC3

Standard deviation 8.0324 1.56982 0.09664

Proportion of Variance 0.9631 0.03678 0.00014

Cumulative Proportion 0.9631 0.99986 1.00000

pca$x

cor(pca$x)

PC1 PC2 PC3

PC1 1.000000e+00 -2.610745e-16 1.769077e-17

PC2 -2.610745e-16 1.000000e+00 2.450821e-16

PC3 1.769077e-17 2.450821e-16 1.000000e+00

pca$rotation

PC1 PC2 PC3

x1cent -0.0001618872 -3.220287e-05 1.000000e+00

x2cent -0.3544056131 -9.350918e-01 -8.748636e-05

x3cent 0.9350917790 -3.544056e-01 1.399665e-04

# Estimated Model:

# PC1 = (-0.0001618872\*x1cent)+(-0.3544056131\*x2cent)+(0.9350917790\*x3cent)

# PC2 = (-3.220287e-05\*x1cent)+(-9.350918e-01\*x2cent)+(-3.544056e-01\*x3cent)

# PC3 = (1.000000e+00\*x1cent)+(-8.748636e-05\*x2cent)+(1.399665e-04\*x3cent)

lm(y~pca$x)

Call:

lm(formula = y ~ pca$x)

Coefficients:

(Intercept) pca$xPC1 pca$xPC2 pca$xPC3

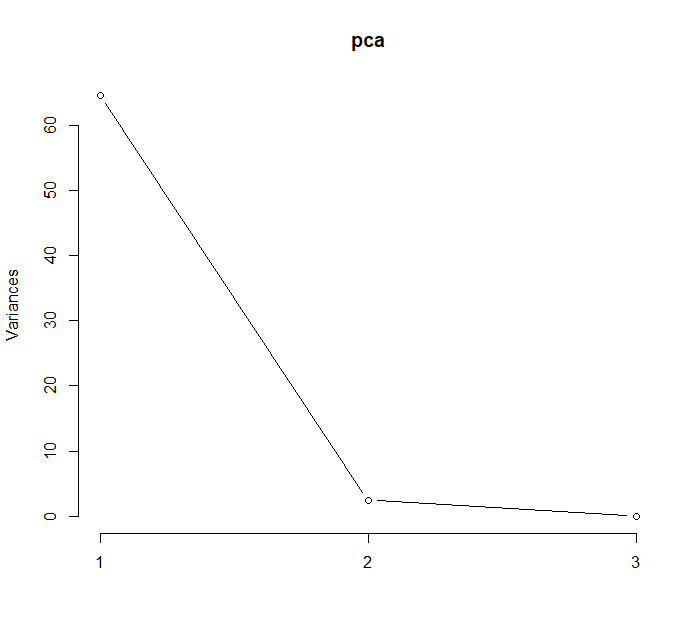
1060.737 -3.761 1.440 -2.603

# Estimated Model:

# y = 1060.737 + (-3.761\*PC1) + (1.440\*PC2) + (-2.603\*PC3)

**1 (d)**

screeplot(pca,type="lines")



lm(y~pca$x[,1])

Call:

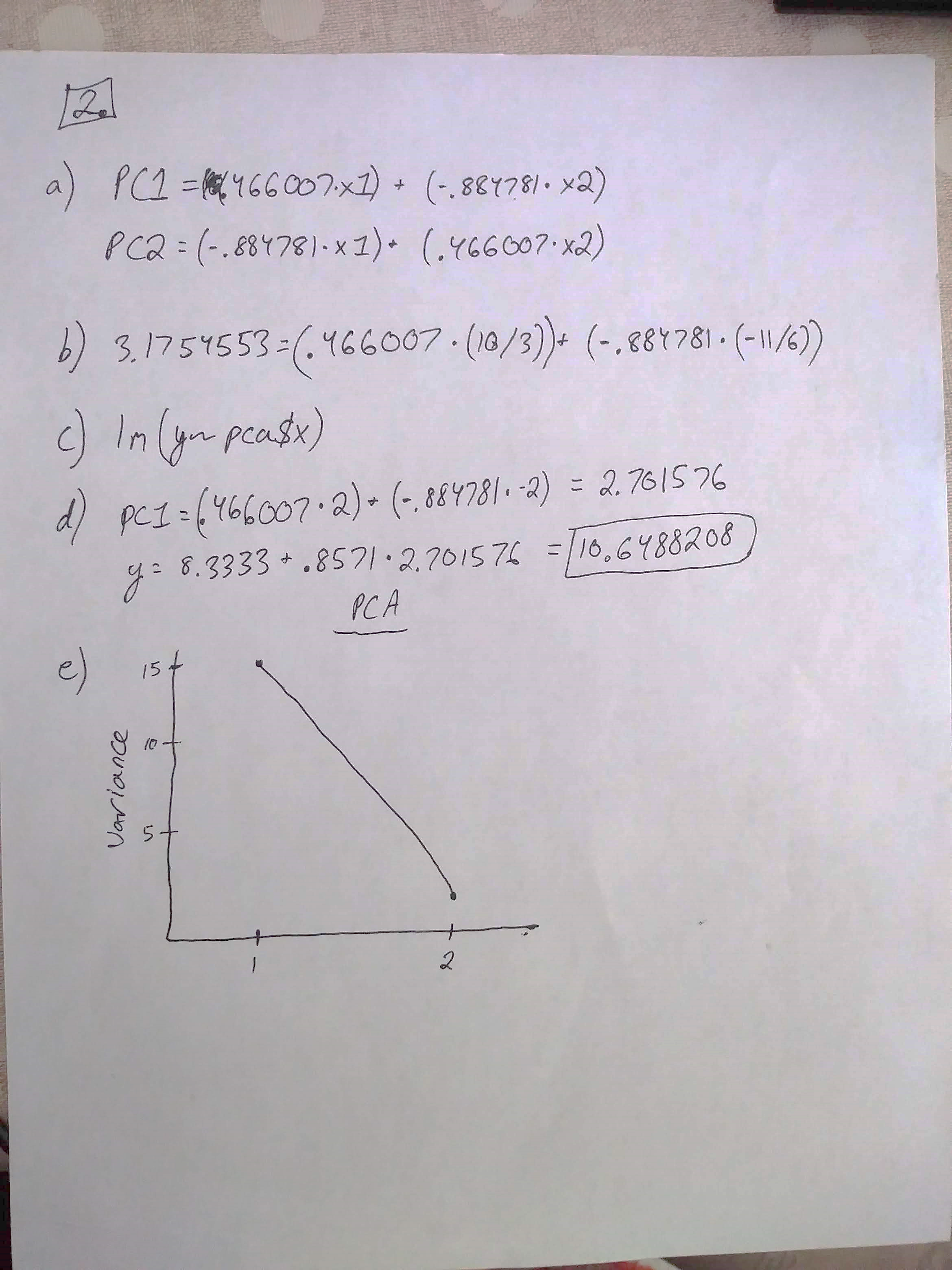
lm(formula = y ~ pca$x[, 1])

Coefficients:

(Intercept) pca$x[, 1]

1060.737 -3.761

### Regression model: y = 1060.737 + (-3.761\*PC1)



**3 (a)**

x1 = sample(1:10,200,replace=TRUE)

x2 = sample(1:10,200,replace=TRUE)

prob = 1/(1+exp(-(2+4\*x1-5\*x2)))

y = rbinom(n=200, size=1, prob=prob)

data = data.frame(x1,x2)

model <- glm(y~.,family=binomial(),data=data)

model

Call: glm(formula = y ~ ., family = binomial(), data = data)

Coefficients:

(Intercept) x1 x2

2.036 3.444 -4.426

Degrees of Freedom: 199 Total (i.e. Null); 197 Residual

Null Deviance: 268.4

Residual Deviance: 28.41 AIC: 34.41

**3 (b)**

Pred <- (fitted(model)>.5)\*1

sum(y!=Pred)/200

0.035

**3 (c)**

Pred <- (fitted(model)>.5)\*1

Pred

The first predicted observation is 0

**3 (d)**

fitted(model)

Probability is 1 of getting a 1