ΔΠΜΣ Βιοστατιστική

και Επιστήμη Δεδομένων Υγείας

Πέτρος Τζαβέλλας

ΑΜ: 7450022400026

1η Εργασία στη Πολυμεταβλητή Στατιστική

Μέρος Α:

1) Αρχικά, πριν προβώ στην εύρεση των πινάκων συσχέτισης των $Y=(Y_1,Y_2)$, $Z=(Z_1,Z_2)$ και $YZ=(Y_1,Z_2)$, θα χρειαστεί να μετασχηματιστεί ο πίνακας συσχέτισης του X σε πίνακα συνδιακύμανσης και για τις δύο περιπτώσεις των διασπορών των X_i , όπου i=1,2,3,4.

Ύστερα, αφού $Y=c^tX$, τότε $V(Y)=c^tΣc$, όπου Σ ο πίνακας συνδιακύμανσης που βρήκαμε παραπάνω, έτσι βάση του παραπάνω τύπου υπολογίζουμε τον πίνακα συνδιακύμανσης του Y και μέσω αυτού υπολογίζουμε τον πίνακα συσχέτισης του Y. Ομοίως κάνουμε για το Z και για τον YZ.

$$R_{y1} = \begin{pmatrix} 1 & -0.54 \\ -0.54 & 1 \end{pmatrix}, \quad R_{y2} = \begin{pmatrix} 1 & -0.64 \\ -0.64 & 1 \end{pmatrix}$$

$$R_{z1} = \begin{pmatrix} 1 & -0.67 \\ -0.67 & 1 \end{pmatrix}, \quad R_{z2} = \begin{pmatrix} 1 & -0.79 \\ -0.79 & 1 \end{pmatrix}$$

$$R_{YZ1} = \begin{pmatrix} 1 & 0.45 \\ 0.45 & 1 \end{pmatrix}, \quad R_{YZ2} = \begin{pmatrix} 1 & 0.0074 \\ 0.0074 & 1 \end{pmatrix}$$

Παρατηρούμαι, όπως είναι λογικό, ότι οι συσχετίσεις διαφέρουν σημαντικά μεταξύ των δύο περιπτώσεων, δηλαδή με ίση διασπορά και με διαφορετική διασπορά μεταξύ των X_i.

2) Προσομοιώνουμε ένα δείγμα μεγέθους 10000 και υπολογίζουμε τον δειγματικό μέσο $\bar{x}$=(1.98,3,4.99,7,99) και δειγματικός πίνακας συνδιακύμανσης τον εξής:

$$S = \begin{pmatrix} 2.0048 & 0.5041 & 1.1241 & 2.5539 \\ 0.5041 & 3.0288 & 1.7122 & 2.3619 \\ 1.1241 & 1.7122 & 3.9811 & 3.1157 \\ 2.5539 & 2.3619 & 3.1157 & 5.0574 \end{pmatrix}$$

Στην συνέχεια, δημιουργούμε τον W=(Y,Z) και βρίσκουμε τον δειγματικό πίνακα συνδυακύμανσής του και κατ' επέκταση τον δειγματικό πίνακα συσχέτισής του.

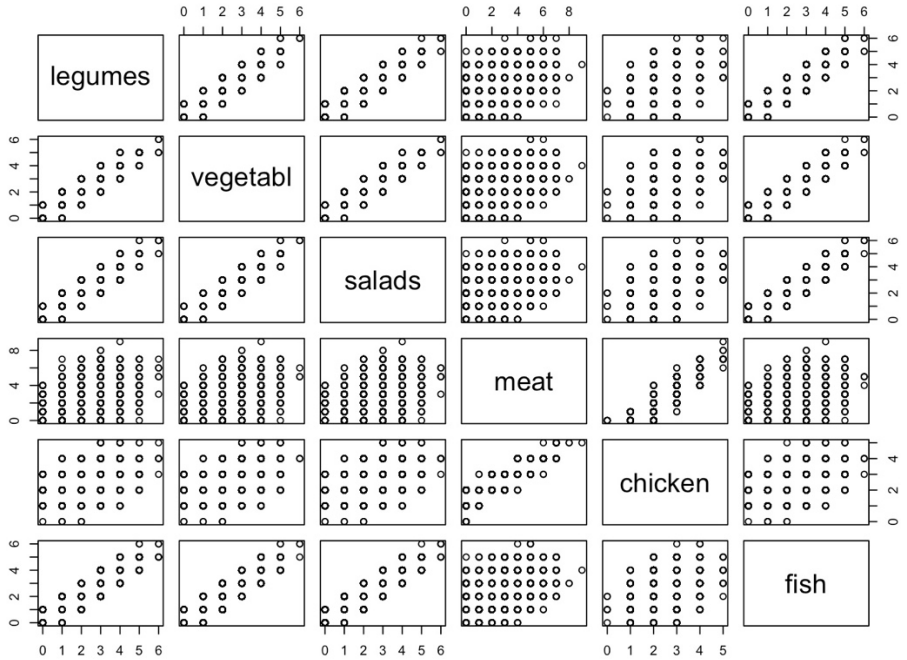
$$R_w = \begin{pmatrix} 1 & -0.63 & 0.25 & 0.004 \\ -0.63 & 1 & -0.4 & 0.38 \\ 0.25 & -0.4 & 1 & -0.79 \\ 0.004 & 0.38 & -0.79 & 1 \end{pmatrix}$$

Παρατηρούμε ότι ο R_W προσεγγίζει αρκετά τους πληθυσμιακούς πίνακες συσχέτισης του ερωτήματος 1, για την δεύτερη περίπτωση όπου η διασπορά δεν είναι σταθερή, κάτι που ήταν αναμενόμενο αφού όσο το δείγμα αυξάνεται προσεγγίζει όλο και περισσότερο της πληθυσμιακές τιμές. Η συσχέτιση μεταξύ Y_1 και Y_2 είναι -0.63 έναντι -0.64, ενώ για το Z_1 και το Z_2 είναι ίδια. Τέλος, μια μικρή διαφορά παρατηρείται στην συσχέτιση του Y_1, Z_2.

Μέρος Β:

1) Παρατηρούμε από τον πίνακα συσχέτισης ότι υπάρχουν αρκετά ικανές συσχετίσεις , κάτι που επιβεβαιώνεται από το KMO που βγαίνει 0,81, άρα μπορούμε να πραγματοποιήσουμε σε ανάλυση σε κύριες συνιστώσες.

$$\begin{pmatrix} 1 & 0.94 & 0.92 & 0.33 & 0.35 & 0.88 \\ 0.94 & 1 & 0.96 & 0.33 & 0.36 & 0.88 \\ 0.92 & 0.96 & 1 & 0.34 & 0.38 & 0.89 \\ 0.33 & 0.33 & 0.34 & 1 & 0.85 & 0.35 \\ 0.35 & 0.36 & 0.38 & 0.85 & 1 & 0.35 \\ 0.88 & 0.88 & 0.89 & 0.35 & 0.35 & 1 \end{pmatrix}$$



2)

a) Οι ιδιοτιμές του πίνακα συνδυακύμανσης είναι οι εξής:

$$\Lambda = \begin{pmatrix} 4.14 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.87 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.14 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.10 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.07 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.03 \end{pmatrix}$$

b) Το ποσοστό της μεταβλητότητας που ερμηνεύει η κάθε συνιστώσα υπολογίζεται ως εξής:

$mp_i = 100 \times \frac{\lambda_i}{\sum \lambda_i}\%$, όπου i, η i-οστη ιδιοτιμή που αντιστοιχεί στην i συνιστώσα.

Κατ' επέκταση, καταλήγουμε στα συγκεκριμένα ποσοστά που ερμηνεύει η κάθε συνιστώσα, σε φθίνουσα σειρά και κάθε ένα αντιστοιχεί σε μία ιδιοτιμή. (65.1%, 29.3%, 2.2%, 1.7%, 1.1%, 0.5%)

c) Τα ιδιοδιανύσματα που αντιστοιχούν στις ιδιοτιμές του πίνακα συνδιακύμανσης είναι τα εξής:

$$E = \begin{pmatrix} 0.404 & 0.299 & -0.157 & 0.255 & -0.773 & 0.246 \\ 0.405 & 0.300 & -0.268 & 0.239 & 0.221 & -0.754 \\ 0.415 & 0.300 & -0.231 & 0.010 & 0.573 & 0.596 \\ 0.532 & -0.754 & 0.208 & 0.316 & 0.058 & 0.029 \\ 0.242 & -0.299 & -0.536 & -0.734 & -0.146 & -0.074 \\ 0.398 & 0.269 & 0.721 & -0.489 & -0.027 & -0.098 \end{pmatrix}$$

Κάθε ιδιοδιάνυσμα αντιστοιχεί σε μία κύρια συνιστώσα. Το πρώτο ιδιοδιάνυσμα αντιστοιχεί στην μεγαλύτερη ιδιοτιμή και εξηγεί το μεγαλύτερο ποσοστό της μεταβλητότητα των αρχικών μεταβλητών. Η δεύτερη συνιστώσα αντιστοιχεί στην δεύτερη μεγαλύτερη ιδιοτιμή και εξηγεί το μεγαλύτερο ποσοστό της εναπομένουσα μεταβλητότητας των αρχικών μεταβλητών, που δεν εξήγησε η πρώτη συνιστώσα και ούτω καθεξής για τις υπόλοιπες .

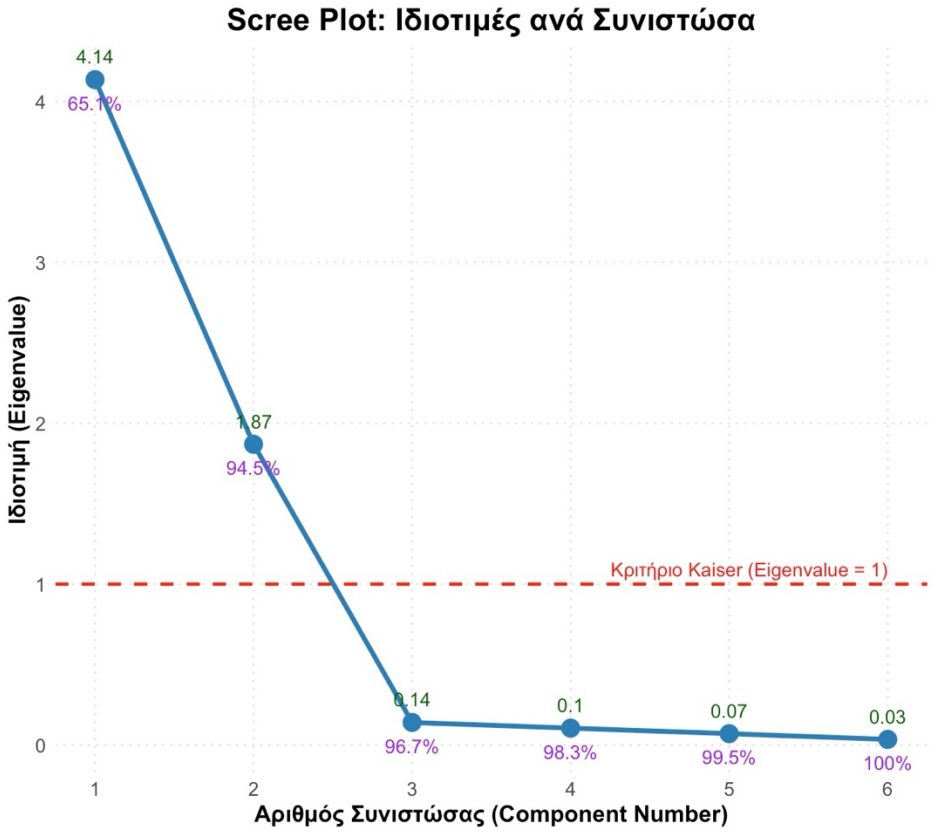
Οι τιμές των ιδιοδιανυσμάτων είναι βάρη (loadings) των αρχικών μεταβλητών, με σκοπό την δημιουργία μια νέας μεταβλητής (κύρια συνιστώσα) που θα είναι ο γραμμικός συνδυασμός των αρχικών μεταβλητών με τα αντίστοιχα βάρη που υποδεικνύει το ιδιοδιάνυσμα.

d) Με οποιοδήποτε εκ των τριών κριτηρίων, καταλήγουμε στο συμπέρασμα να κρατήσουμε 2 κύριες συνιστώσες.

Αρχικά με κριτήριο Kaizer, κρατάμε όσες συνιστώσες είναι μεγαλύτερες του 1 και κατ' επέκταση κρατάμε 2 συνιστώσες, αφού όπως παρατηρούμαι στο ερώτημα a) μόνο οι πρώτες δύο ξεπερνάνε αυτό το όριο.

Από τον κανόνα ότι το ποσοστό συνολικής διακύμανσης που εξηγούν οι κύριες συνιστώσες πρέπει να είναι μεγαλύτερο του 80% καταλήγουμε σε 2 κύριες συνιστώσες, αφού οι πρώτες 2 συνιστώσες εξηγούν 94.4%, ποσοστό αρκετά μεγαλύτερο του 80%.

Τέλος, από το scree plot που φαίνεται παρακάτω καταλήγουμε στο ίδιο συμπέρασμα.

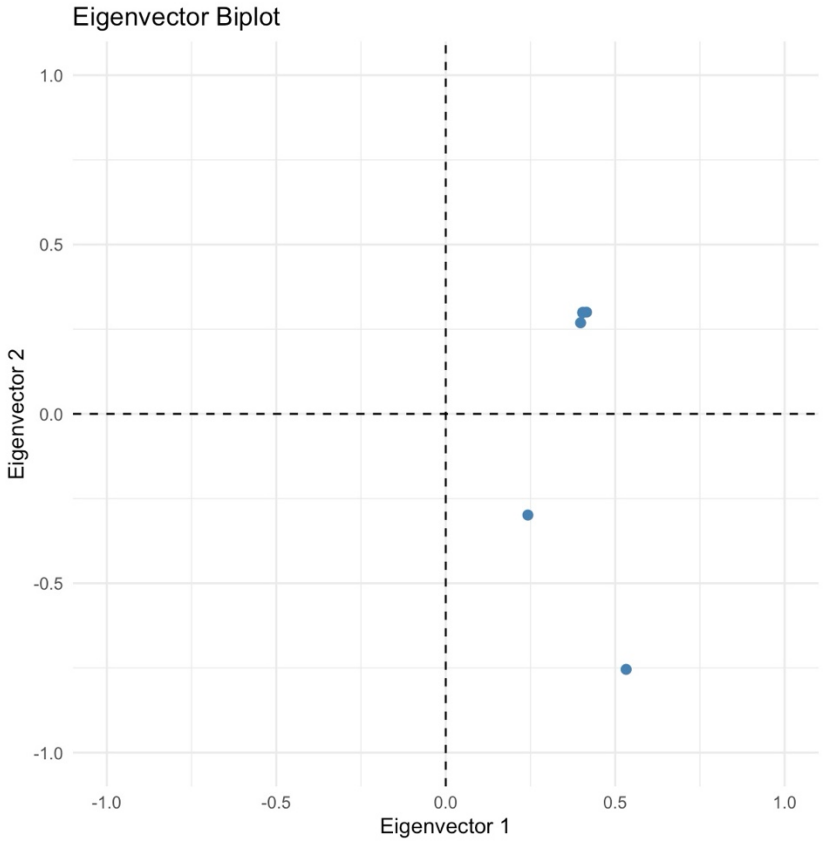


e)

Ερμηνεία πρώτης συνιστώσας: Η πρώτη κύρια συνιστώσα αντιπροσωπεύει έναν σταθμισμένο μέσο όρο όλων των ομάδων τροφίμων, με θετικά βάρη (loadings) σε όλες τις μεταβλητές. Αυτό

υποδηλώνει, ότι εκφράζει ένα μοτίβο γενικά αυξημένης κατανάλωσης τροφίμων, ανεξαρτήτως είδους. Υψηλές τιμές της, αντιστοιχούν σε άτομα που καταναλώνουν συχνότερα όλες τις ομάδες τροφών, γεγονός που ενδεχομένως αντικατοπτρίζει μια πιο ισορροπημένη διατροφή.

Ερμηνεία δεύτερης συνιστώσας: Δείχνει την τάση των ασθενών στην διατροφή που κάνουν. Υψηλά σκορ στην δεύτερη συνιστώσα δείχνουν μια pescatarian διατροφή, αντίθετα χαμηλά σκορ δείχνουν μια πιο κρεατοφαγική διατροφή



f) Από τον παρακάτω πίνακα παρατηρούμε ότι η πρώτη συνιστώσα δεν σχετίζεται με την ενδονοσοκομειακή θνητότητα των ασθενών της μελέτης. Αντίθετα, η δεύτερη συνιστώσα φαίνεται να σχετίζεται με την ενδονοσοκομειακή θνητότητα έπειτα από οξύ έμφραγμα του μυοκαρδίου.

Συμπεραίνουμε ότι, για κάθε μονάδα αύξησης της δεύτερης συνιστώσας η πιθανότητα ενδονοσοκομειακού θανάτου έπειτα από οξύ έμφραγμα του μυοκαρδίου μειώνεται κατά 15%.

A/A	Estimate	Standard Error	Value	P-value
Intercept	-0.93	0.07	-13.5	2e-16
1st Component	0.0092	0.01	0.84	0.4
2nd Component	-0.16	0.02	-10.07	2e-16

Παράρτημα:

```
library(MASS)
library(ggplot2)
library(psych)

####       PART 1       ####

##  1.  ##

mu <- c(2, 3, 5, 8)

R <- matrix(c(
  1,   0.2, 0.4, 0.8,
  0.2, 1,   0.5, 0.6,
  0.4, 0.5, 1,   0.7,
  0.8, 0.6, 0.7, 1
), nrow=4, byrow=TRUE)
```

i)

```
var1=rep(sqrt(2),4)
D1=diag(var1)
Sigma1=D1%*%R%*%D1

## Correlation Y ##

cy1=rbind(c(1,-1,0,-1),c(0,0,1,0))
Sy1=cy1%*%Sigma1%*%t(cy1)
Vy1=sqrt(diag(Sy1))
Ry1=Sy1/(Vy1%*%t(Vy1))


## Correlation Z ##

cz1=rbind(c(-1,0,0,0),c(0,-1,0,1))
Sz1=cz1%*%Sigma1%*%t(cz1)
Vz1=sqrt(diag(Sz1))
Rz1=Sz1/(Vz1%*%t(Vz1))

## Correlation Y1 & Z2 ##

cyz1=rbind(c(1,-1,0,-1),c(0,-1,0,1))
Syz1=cyz1%*%Sigma1%*%t(cyz1)
Vyz1=sqrt(diag(Syz1))
Ryz1=Syz1/(Vyz1%*%t(Vyz1))
```

ii)

```
var2=sqrt(c(2,3,4,5))
D2=diag(var2)
Sigma2=D2%*%R%*%D2

## Correlation Y ##

cy2=rbind(c(1,-1,0,-1),c(0,0,1,0))
Sy2=cy2%*%Sigma2%*%t(cy2)
Vy2=sqrt(diag(Sy2))
Ry2=Sy2/(Vy2%*%t(Vy2))

## Correlation Z ##

cz2=rbind(c(-1,0,0,0),c(0,-1,0,1))
Sz2=cz2%*%Sigma2%*%t(cz2)
Vz2=sqrt(diag(Sz2))
Rz2=Sz2/(Vz2%*%t(Vz2))

## Correlation Y1 & Z2 ##

cyz2=rbind(c(1,-1,0,-1),c(0,-1,0,1))
Syz2=cyz2%*%Sigma2%*%t(cyz2)
Vyz2=sqrt(diag(Syz2))
Ryz2=Syz2/(Vyz2%*%t(Vyz2))
```

```
## 2. ##

## Setting Parameters for multivariate Normal ##

n=10000
Sigma=Sigma2
mu

## Setting seed for reproduction purposes ##

set.seed(2025)

## Taking sample of 10000 multivariate Normals ##

X=mvrnorm(n, mu, Sigma)

## Calculating sample means and covariances of X ##

X_bar=apply(X,2,mean)
X_cov=cov(X)

## Computing y1,y2,z1,z2 and creating w ##

y1=X[,1]-X[,2]-X[,4]
y2=X[,3]
z1=-X[,1]
```

```
z2=X[,4]-X[,2]

w=cbind(y1,y2,z1,z2)


## Calculating sample means, covariances and correlation of w ##


w_bar=apply(w,2,mean)

w_cov=cov(w)

Varw=sqrt(diag(w_cov))

Rw=w_cov/(Varw%*%t(Varw))




##------------------------------------------------------------------------------##


####       PART 2       ####

setwd("//Users//petros//Library//Mobile Documents//com~apple~CloudDocs//MSc Biostatistics//Πολυμεταβλητή Στατιστική//Assignments//Assignment 1")

diet=read.csv("dietStudy.csv")

deaths=diet$death

diet=diet[,-7]

dcov=cov(diet)



##  1. ##


round(cor(diet),2)

pairs(diet)

KMO(diet)
```

```
##  2.  ##


## a) ##


eigd <- eigen(dcov)


## b) ##


eigd$cp = eigd$values/sum(eigd$values)

eigd$vectors=-(eigd$vectors)

d_eig=data.frame(it =
1:ncol(diet),values=eigd$values,cp=eigd$cp,vectors=eigd$vectors)

d_eig$ceig <- cumsum(d_eig$cp)


## d) ##

d_eig$ceig=d_eig$ceig*100

head(d_eig[d_eig$ceig>80,],1)

d_eig[d_eig$values>1,]

k=length(d_eig$values)


ggplot(d_eig, aes(x = it, y = values)) +
 geom_line(color = "#2C7FB8", linewidth = 1.2) + # Γραμμή
 geom_point(color = "#2C7FB8", size = 4, shape = 19) + # Σημεία
 labs(
  title = "Scree Plot: Ιδιοτιμές ανά Συνιστώσα",
  x = "Αριθμός Συνιστώσας (Component Number)",
  y = "Ιδιοτιμή (Eigenvalue)"
```

```
 ) +
 # Προσθήκη οριζόντιας γραμμής για το Κριτήριο Kaiser
 geom_hline(yintercept = 1, linetype = "dashed", color = "red", linewidth = 0.8) +
 annotate("text", x = max(d_eig$it), y = 1.05, label = "Κριτήριο Kaiser (Eigenvalue = 1)",
       color = "red", hjust = 1, vjust = 0, size = 3.5) +
 # Προσθήκη ετικετών με τις ακριβείς τιμές των ιδιοτιμών (πάνω από τα σημεία)
 geom_text(aes(label = round(values, 2)), # Ετικέτα: στρογγυλοποιημένη τιμή ιδιοτιμής
       vjust = -1.2, # Μετακίνηση ετικέτας πάνω από το σημείο
       hjust = 0.5,
       size = 3.5,
       color = "darkgreen") +
 # ΝΕΟ: Προσθήκη ετικετών με την αθροιστική διακύμανση (κάτω από τα σημεία)
 geom_text(aes(label = paste0(round(ceig, 1), "%")), # Ετικέτα: στρογγυλοποιημένο ceig με "%"
       vjust = 2.2, # Μετακίνηση ετικέτας κάτω από το σημείο (μεγαλύτερη τιμή vjust = πιο κάτω)
       hjust = 0.5,
       size = 3.5,
       color = "purple") + # Χρησιμοποιούμε διαφορετικό χρώμα για διαφοροποίηση
 # Εξασφάλιση ότι ο x-άξονας δείχνει ακέραιους αριθμούς
 scale_x_continuous(breaks = 1:k) +
 # Επιλογή ενός καθαρού, μινιμαλιστικού θέματος
 theme_minimal() +
 # Προσαρμογή του θέματος για καλύτερη εμφάνιση
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 16), # Κεντράρισμα και έντονη γραμματοσειρά τίτλου
  axis.title = element_text(size = 12, face = "bold"),
  axis.text = element_text(size = 10),
```

```
  panel.grid.major = element_line(color = "gray90", linetype = "dotted"), # Κύριες γραμμές πλέγματος
  panel.grid.minor = element_blank() # Αφαίρεση δευτερευόντων γραμμών πλέγματος
 )

## e ##

ggplot(d_eig, aes(x = vectors.1, y = vectors.2)) +
 geom_point(color = "steelblue", size = 2) +
 geom_hline(yintercept = 0, color = "black", linetype = "dashed") +  # y = 0 line
 geom_vline(xintercept = 0, color = "black", linetype = "dashed") +  # x = 0 line
 coord_fixed(xlim = c(-1, 1), ylim = c(-1, 1)) +  # fixed ratio and limits
 labs(title = "Eigenvector Biplot",
     x = "Eigenvector 1",
     y = "Eigenvector 2") +
 theme_minimal()

## f ##

diet[,c("Comp1","Comp2")] <- as.matrix(diet) %*% (eigd$vectors[,1:2])
diet$death=deaths
glm_log=glm(death~Comp1+Comp2,family = binomial(link="logit"),data = diet)
summary(glm_log)
```