Practica III

Robinson Aldair Cuayal

2023-03-11

Table of Contents

# Punto 1

El ejercicio en cuestión implica la realización de un análisis de regresión para determinar si existe una relación entre el nivel de plomo en el suelo y la categoría del individuo, y si el nivel de plomo en el suelo es un factor significativo para predecir los niveles altos de plomo en la sangre de los niños. Se proporcionan la hipótesis nula y la alternativa, así como los valores de p, la ecuación de regresión y las estadísticas relevantes. El objetivo principal es predecir los niveles altos de plomo en la sangre de los niños. Como primer paso, se va a almacenar la informacion de las tablas en un excel para luego ser cargado y convertido a un dataframe en R.

library(readxl)

## Warning: package 'readxl' was built under R version 4.2.3

datos = data.frame(read\_excel("dataset\_ejer1.xlsx"))  
dim(datos)

## [1] 139 2

head(datos)

## Categoria\_del\_Individuo Nivel\_plomo\_suelo  
## 1 1 1290  
## 2 0 90  
## 3 1 894  
## 4 0 193  
## 5 1 1410  
## 6 1 410

Con la carga de los datos, como segundo paso se implementa el modelo de regresion logistica. Cabe resaltar que es importante especificar el tipo de enlace en la modelación logística ya que afecta directamente la interpretación de los coeficientes del modelo y los resultados obtenidos.

# Crear la regresión logística  
modelo\_logistico = glm(  
 Categoria\_del\_Individuo ~ Nivel\_plomo\_suelo,  
 data = datos,  
 family = binomial(link = "logit")  
)  
#modelo general  
summary(modelo\_logistico)

##   
## Call:  
## glm(formula = Categoria\_del\_Individuo ~ Nivel\_plomo\_suelo, family = binomial(link = "logit"),   
## data = datos)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.27743 -0.80117 0.05973 0.85069 1.80313   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.516206 0.339557 -4.465 8.0e-06 \*\*\*  
## Nivel\_plomo\_suelo 0.002743 0.000544 5.042 4.6e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 191.82 on 138 degrees of freedom  
## Residual deviance: 140.24 on 137 degrees of freedom  
## AIC: 144.24  
##   
## Number of Fisher Scoring iterations: 6

# Obtener la razón de grados de probabilidad  
coeficientes = coef(modelo\_logistico)  
razon\_encontrada = exp(coeficientes["Nivel\_plomo\_suelo"])  
razon\_encontrada

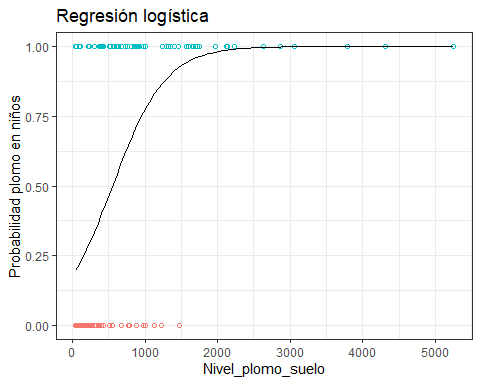
## Nivel\_plomo\_suelo   
## 1.002747

# Compara la razón de grados de probabilidad contra la que obtuvieron los autores (14.25)  
razon\_autores = 14.25  
# Realizar la prueba de significancia  
summary(modelo\_logistico)$coefficients["Nivel\_plomo\_suelo", "Pr(>|z|)"]

## [1] 4.601328e-07

La razón de grados de probabilidad obtenida es de 1.002, lo que indica que por cada diez veces que aumenta el nivel de plomo en la tierra, la proporción relativa de casos con respecto a los grupos de control aumenta en un 0.1. Este resultado difiere del obtenido por los autores del estudio (14.25). El valor de p obtenido es , lo que indica que el nivel de plomo en la tierra es un factor significativo para predecir los niveles altos de plomo en la sangre de los niños. Las grafica del modelo es el siguiente

library(ggplot2)  
ggplot(data = datos, aes(x = Nivel\_plomo\_suelo, y = Categoria\_del\_Individuo)) +  
 geom\_point(aes(color = as.factor(Categoria\_del\_Individuo)), shape = 1) +  
 stat\_function(fun = function(x){predict(modelo\_logistico,  
 newdata = data.frame(Nivel\_plomo\_suelo = x),  
 type = "response")}) +  
 theme\_bw() +  
 labs(title = "Regresión logística",  
 y = "Probabilidad plomo en niños") +  
 theme(legend.position = "none")



## Preguntas

1. La variable dependiente es la “Categoria\_del\_Individuo”.
2. La variable independiente es la “Nivel\_plomo\_suelo”.
3. Hipótesis nula: No hay relación entre el nivel de plomo en el suelo y la categoría del individuo. Hipótesis alternativa: Existe una relación entre el nivel de plomo en el suelo y la categoría del individuo.
4. La hipótesis nula es rechazada debido a que el valor de p obtenido es menor al nivel de significancia del 0.05 establecido. Esto indica que hay suficiente evidencia para concluir que el nivel de plomo en el suelo es un factor significativo para predecir los niveles altos de plomo en la sangre de los niños.
5. El objetivo más relevante es predecir los niveles altos de plomo en la sangre de los niños, ya que esto puede tener graves consecuencias en su salud. Aunque la estimación de parámetros es importante para comprender la relación entre las variables, la predicción tiene una mayor importancia práctica.
6. La población muestreada son niños expuestos al plomo en la ciudad donde se realizó el estudio.
7. La población objetivo son todos los niños expuestos al plomo, independientemente de su ubicación geográfica.
8. La variable “Nivel\_plomo\_suelo” está relacionada directamente con la “Categoria\_del\_Individuo”, ya que aumentos en la primera variable son predictores de una mayor proporción de casos en la segunda variable, al examinar la significación estadística de los coeficientes de la regresión y de las variables incluidas al ser significativo y positivo. Los datos no tienen una alta correlacion segun el Nab pero al tener los resultado del modelo de regresion logistica hay que tener en cuenta el primer enunciado.

r = cor(datos)  
nab = 1-det(r)  
nab

## [1] 0.2290155

1. La ecuación de regresión es: , donde “y” es la “Categoria\_del\_Individuo”, “x” es la “Nivel\_plomo\_suelo”, b0 es el intercepto y b1 es la pendiente. En este caso, la ecuación sería: .
2. Para realizar otro analisis adicionando valores puede ser el del sexo que pueden tomar como masculino (1) y femenino (0).
3. La variable “Categoria\_del\_Individuo” es categórica binaria (0 para control y 1 para caso), mientras que la variable “Nivel\_plomo\_suelo” es cuantitativa. La razón de grados de probabilidad (también conocido como odds ratio) es una medida de la asociación entre dos eventos. La variable sexo tambien es categórica binaria. diferentes valores de la variable independiente.
4. Con base en el punto (j) donde se le puede dar valores de nuestra parte a la variable de sexo de los niños, se diseña el modelo modelo logistico para el siguiente analisis estadistico.

#codigo modelo 2 aqui

# Punto 2

En primer lugar, se realiza la carga de los datos que se almacenaron previamente en una hoja excel, para luego convertir a un dataframe:

library(readxl)  
olmos <- data.frame(read\_excel("dataset\_ejer\_2.xlsx"))  
dim(olmos)

## [1] 30 7

head(olmos)

## X1 X2 X3 X4 X5 X6 X7  
## 1 0.530 0.420 0.135 0.6770 0.2565 0.1415 0.210  
## 2 0.530 0.415 0.150 0.7775 0.2370 0.1415 0.330  
## 3 0.545 0.425 0.125 0.7680 0.2940 0.1495 0.260  
## 4 0.550 0.440 0.150 0.8945 0.3145 0.1510 0.320  
## 5 0.525 0.380 0.140 0.6065 0.1940 0.1475 0.210  
## 6 0.535 0.405 0.145 0.6845 0.2725 0.1710 0.205

## Preguntas

1. Realizar el analisis de componenetes principales (PCA) completo con gráficos de cargas.

Antes de realizar un analisis por PCA, hay que tener en cuenta que debe haber una alta correlacion entre los datos, porque de lo contrario no tiene sentido realizar este analisis.

r = det(cor(olmos))  
nab = 1-r  
nab

## [1] 0.9999999

Como se puede observar hay una alta correlacion de los datos ya que el . Ya con esta prueba es factible hacer este analisis de PCA. Se comienza con la creacion del modelo PCA, donde esta informacion relevante como varianza y correlacion.

# Análisis de componentes principales  
acp = princomp(olmos, cor = TRUE)  
summary(acp)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 2.5273682 0.53058381 0.44469355 0.28934876 0.161625395  
## Proportion of Variance 0.9125128 0.04021703 0.02825034 0.01196039 0.003731824  
## Cumulative Proportion 0.9125128 0.95272986 0.98098020 0.99294059 0.996672411  
## Comp.6 Comp.7  
## Standard deviation 0.136607246 0.0680557282  
## Proportion of Variance 0.002665934 0.0006616546  
## Cumulative Proportion 0.999338345 1.0000000000

#grafico de cargas de las variables  
vector\_propios = acp$loadings  
#cambio de signo  
acp$loadings[,2] = -acp$loadings[,2]  
acp$scores[,2] = -acp$scores[,2]  
#correlacion entre las componentes y las variables  
desviacion\_stand = acp$sdev  
correlacion = sweep(vector\_propios[1:7,1:7],2, desviacion\_stand, "\*" )  
round(correlacion,3)

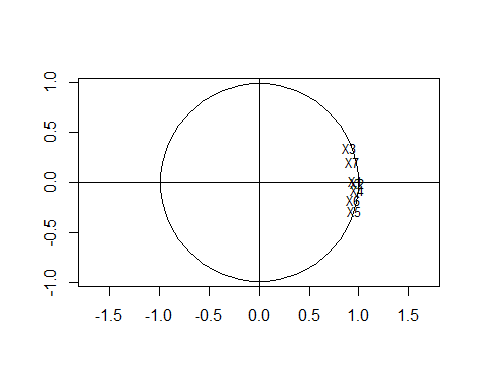
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7  
## X1 0.968 0.011 0.050 0.240 0.045 0.037 0.004  
## X2 0.985 -0.006 0.103 0.055 -0.099 -0.084 0.006  
## X3 0.903 0.348 -0.246 -0.026 -0.043 0.032 -0.004  
## X4 0.989 -0.079 0.093 -0.057 0.009 0.012 -0.058  
## X5 0.951 -0.286 -0.008 -0.072 -0.051 0.072 0.024  
## X6 0.948 -0.172 -0.246 -0.031 0.076 -0.062 0.007  
## X7 0.940 0.206 0.235 -0.115 0.063 -0.005 0.023

Se realiza el grafico de cargas de en donde para identificar cuál variable tiene más contribución en el gráfico de cargas circulares, se debe observar la distancia de cada círculo al origen central del gráfico. Por lo cual todas las variables tienen un buen aporte de informacion.

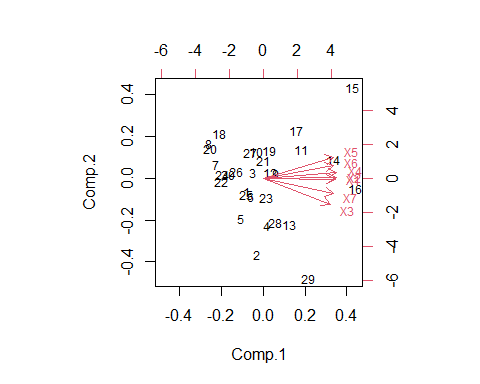
#grafico de cargas de las variables  
library(MASS)

## Warning: package 'MASS' was built under R version 4.2.3

eqscplot(correlacion[,1:2], xlim=c(-1,1), ylim=c(-1,1), type = "n")  
abline(h=0,v=0)  
text(correlacion[,1:2],labels=colnames(olmos),cex=0.8)  
symbols(0,0,circles=1, inches=FALSE, add=TRUE)



biplot(acp, cex=0.75)



1. Cuales variables presentan más similitud.

Segun la tabla, se observa que la variable con mayor similitud es X2 (diametro) y X4 (peso total), con un valor de correlación de 97%. Tambien con una correlacion del 96% se tiene relacion las siguientes parejas (X1,X2), (X4,X5)

#Variable con mayor similitud  
cor(olmos)

## X1 X2 X3 X4 X5 X6 X7  
## X1 1.0000000 0.9635989 0.8580583 0.9480612 0.9000948 0.8975352 0.8992660  
## X2 0.9635989 1.0000000 0.8614770 0.9784638 0.9323299 0.9053611 0.9370721  
## X3 0.8580583 0.8614770 1.0000000 0.8441517 0.7671737 0.8522890 0.8631614  
## X4 0.9480612 0.9784638 0.8441517 1.0000000 0.9656858 0.9301455 0.9413077  
## X5 0.9000948 0.9323299 0.7671737 0.9656858 1.0000000 0.9470957 0.8387932  
## X6 0.8975352 0.9053611 0.8522890 0.9301455 0.9470957 1.0000000 0.8076783  
## X7 0.8992660 0.9370721 0.8631614 0.9413077 0.8387932 0.8076783 1.0000000

1. Cuantos componentes explican la mayor variabilidad de los datos.

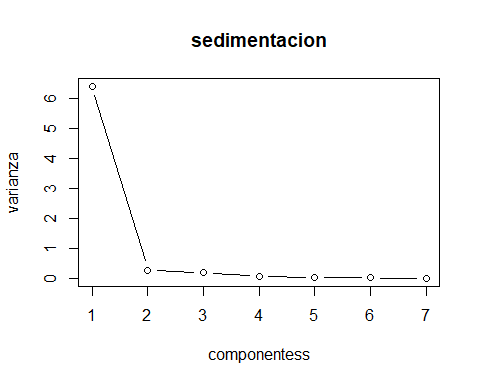
Para conocer cuántos componentes explican la mayor variabilidad de los datos, se usa la función summary() sobre el objeto pca. La salida muestra que el componente 1 explica el 91.25% de la variabilidad. Tambien en el grafico de sedimentacion se observo que era la unica componente al superar el 1, por lo cual solo se escoge esta componente.

summary(acp)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 2.5273682 0.53058381 0.44469355 0.28934876 0.161625395  
## Proportion of Variance 0.9125128 0.04021703 0.02825034 0.01196039 0.003731824  
## Cumulative Proportion 0.9125128 0.95272986 0.98098020 0.99294059 0.996672411  
## Comp.6 Comp.7  
## Standard deviation 0.136607246 0.0680557282  
## Proportion of Variance 0.002665934 0.0006616546  
## Cumulative Proportion 0.999338345 1.0000000000

Se realiza el grafico de sedimentacion en donde se escoge las variables que tengan una varianza mayor a 1.

#calculo de los valores propios  
val\_propios = acp$sdev^2  
plot(1:7, val\_propios, type= "b", xlab="componentess", ylab="varianza",   
main = "sedimentacion")



1. Que grupos de individuos presentan más similitud.

Para conocer los grupos de individuos que presentan mayor similitud se realiza la grafica de dispersion de los individuos y las variables. Segun el grafico se encuentra dos grupos divididos por una linea vertical en el punto 0. El gráfico muestra que las dos observaciones están relativamente cercanas entre sí.

library(psych)

## Warning: package 'psych' was built under R version 4.2.3

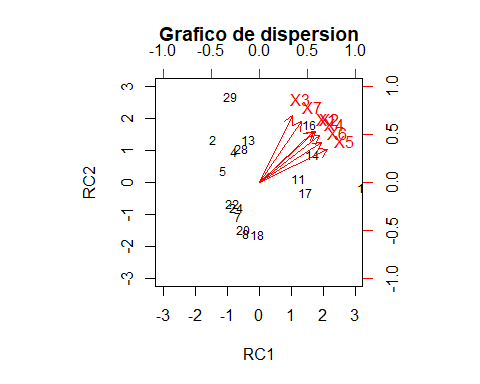
##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

acp\_varimax = principal(olmos, nfactors = 2, rotate = "varimax",scores = TRUE)  
round(cor(olmos, acp\_varimax$scores[,1:2]),3)#tabla de comunidades

## RC1 RC2  
## X1 0.705 0.664  
## X2 0.728 0.663  
## X3 0.428 0.868  
## X4 0.781 0.612  
## X5 0.893 0.434  
## X6 0.814 0.517  
## X7 0.552 0.789

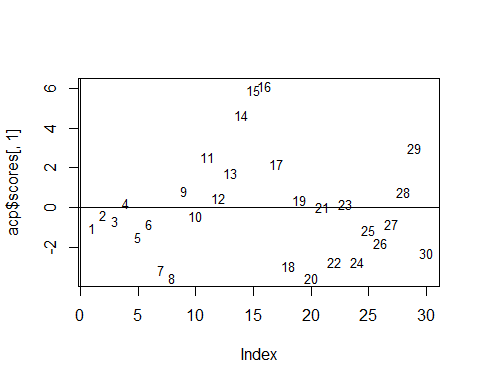
biplot(acp\_varimax, labels = rownames(olmos), cex= 0.75, main = "Grafico de dispersion")



1. Realizar un análisis de regresión lineal del componente 1 y sus respectivas variables de carga.

El siguiente grafico ayuda a ver el aporte que tienen cada individuo al componenetes 1.

plot(acp$scores[,1], type= "n",)  
abline(h=0, v= 0)  
text(acp$scores[,1],   
labels= rownames(olmos), cex=0.8)



Para realizar el análisis de regresión lineal del componente 1 y su respectiva variables de carga, se crea un data frame con las dos columnas.

#para la variable X1  
datos\_regresion = data.frame(acp$scores[,1], olmos$X1)  
names(datos\_regresion) = c('cargasX1', 'datosX1')  
model\_X1 = lm(datos\_regresion$datosX1~datos\_regresion$cargasX1)   
summary(model\_X1)

##   
## Call:  
## lm(formula = datos\_regresion$datosX1 ~ datos\_regresion$cargasX1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.056653 -0.010549 0.002421 0.009731 0.023188   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.541667 0.003133 172.91 <2e-16 \*\*\*  
## datos\_regresion$cargasX1 0.025209 0.001239 20.34 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01716 on 28 degrees of freedom  
## Multiple R-squared: 0.9366, Adjusted R-squared: 0.9343   
## F-statistic: 413.6 on 1 and 28 DF, p-value: < 2.2e-16

#para la varible X2  
datos\_regresion2 = data.frame(acp$scores[,1], olmos$X2)  
names(datos\_regresion2) = c('cargasX1', 'datosX2')  
model\_X2 = lm(datos\_regresion2$datosX2~datos\_regresion2$cargasX1)   
summary(model\_X2)

##   
## Call:  
## lm(formula = datos\_regresion2$datosX2 ~ datos\_regresion2$cargasX1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.015248 -0.006064 -0.002282 0.006445 0.029078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4265000 0.0018978 224.7 <2e-16 \*\*\*  
## datos\_regresion2$cargasX1 0.0223745 0.0007509 29.8 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01039 on 28 degrees of freedom  
## Multiple R-squared: 0.9694, Adjusted R-squared: 0.9683   
## F-statistic: 887.9 on 1 and 28 DF, p-value: < 2.2e-16

Como se puede ver hay una alta correlacion entre los datos y el nuevo componente 1 PCA, el cual explica el 91% de los datos, por lo cual cuando se realiza regresion lineal con cada una de las variables de los datos da una alta correlacion y con un mayor al 90%. Solo se hizo modelos lm() hasta la variable 2 los modelos. En conclusión, el análisis de regresión lineal del primer componente principal y sus respectivas variables de carga permite explicar en gran medida la variabilidad de los datos y evidencia la influencia de la variables X1 en la formación del componente. La carga de una variable en un componente principal indica cuánto contribuye esa variable a la varianza explicada por ese componente. Al realizar la regresión lineal, puedes analizar cómo la variable original está relacionada con el componente principal y si existe una relación lineal significativa entre ellas.

# Punto 3

library(readxl)  
datos <- data.frame(read\_excel("dataset\_ejer\_3.xlsx"))  
dim(datos)

## [1] 10 7

# Punto 4

cat('hola')

## hola

# Punto 5

cat('hola')

## hola