BACS HW14

106070020

2021年5月28日

Question 1

(a)Let’s try computing the direct effects first:

(i)Model 1: Regress log.weight. over log.cylinders. only and report the coefficient

```
#Q1(a)(i)
cars <- read.table("C:/Users/eva/Desktop/作業 上課資料(清大)/大四下/BACS/HW11 BACS/auto-data.txt", header=F, na.strings
 = "?")
names(cars) <- c("mpg", "cylinders", "displacement", "horsepower", "weight",
                 "acceleration", "model_year", "origin", "car_name")
cars_log <- with(cars, data.frame(log(mpg), log(cylinders), log(displacement), log(horsepower), log(weight), log(accel
eration), model_year, origin))
cars_log<-na.omit(cars_log)
wc<-lm(log.weight.~log.cylinders., data=cars_log)
summary(wc)
```

```
##
## Call:
## lm(formula = log.weight. ~ log.cylinders., data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35409 -0.09030 -0.00169  0.09271  0.40488
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.60059    0.03710  177.92   <2e-16 ***
## log.cylinders.  0.82187    0.02208   37.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1319 on 390 degrees of freedom
## Multiple R-squared:  0.7804, Adjusted R-squared:  0.7798
## F-statistic:  1386 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
round(summary(wc)$coefficients,2)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.60       0.04  177.92        0
## log.cylinders.     0.82       0.02   37.23        0
```

> The P-value<0.05, so the number of cylinders has a significant direct effect on weight.

(ii)Model 2: Regress log.mpg. over log.weight. and all control variables and report the coefficient

```
#(a)(ii)
mw<-lm(log.mpg.~log.weight.+ log.acceleration.+ model_year+factor(origin), data=cars_log)
summary(mw)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + model_year +
##     factor(origin), data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38259 -0.07054  0.00401  0.06696  0.39798
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.410974   0.316806  23.393  < 2e-16 ***
## log.weight.       -0.875499   0.029086 -30.101  < 2e-16 ***
## log.acceleration.  0.054377   0.037132   1.464  0.14389
## model_year         0.032787   0.001731  18.937  < 2e-16 ***
## factor(origin)2    0.056111   0.018241   3.076  0.00225 **
## factor(origin)3    0.031937   0.018506   1.726  0.08519 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1163 on 386 degrees of freedom
## Multiple R-squared:  0.8845, Adjusted R-squared:  0.883
## F-statistic: 591.1 on 5 and 386 DF,  p-value: < 2.2e-16
```

```
round(summary(mw)$coefficients,2)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.41       0.32   23.39     0.00
## log.weight.          -0.88       0.03  -30.10     0.00
## log.acceleration.     0.05       0.04    1.46     0.14
## model_year            0.03       0.00   18.94     0.00
## factor(origin)2       0.06       0.02    3.08     0.00
## factor(origin)3       0.03       0.02    1.73     0.09
```

> The P-value<0.05, so the number of weight has a significant direct effect on mpg.

(b)What is the indirect effect of cylinders on mpg?

```
#Q1(b)
indirect_effect<-round(summary(wc)$coefficients,2)[2,1]*round(summary(mw)$coefficients,2)[2,1]
indirect_effect
```

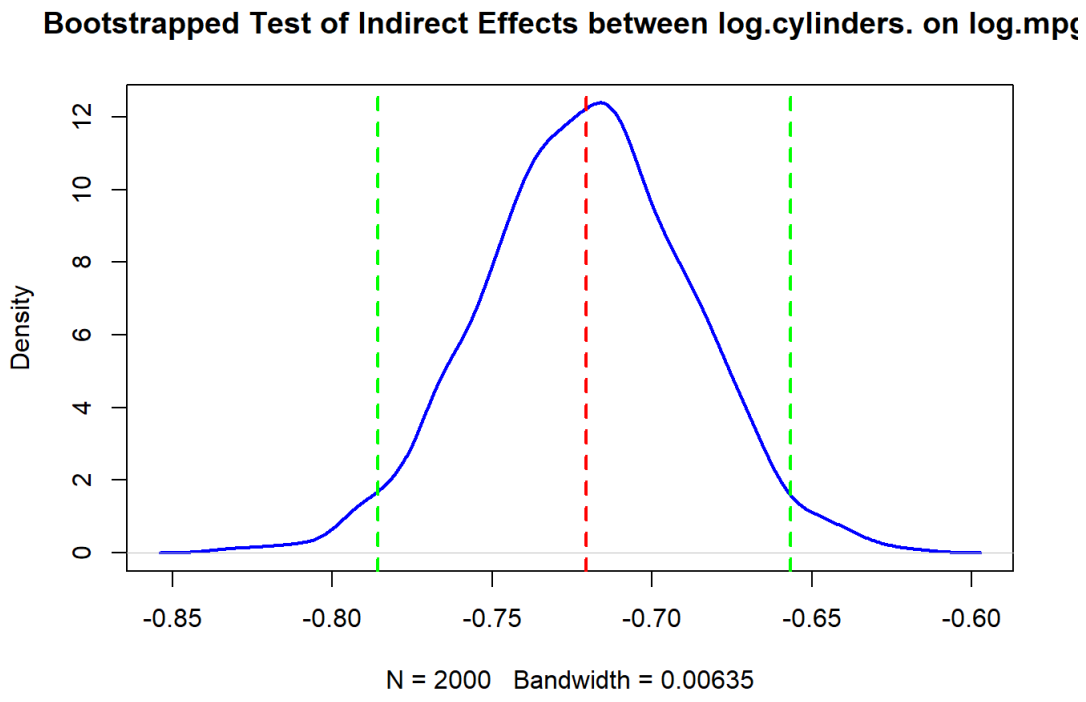
```
## [1] -0.7216
```

(c)Let's bootstrap for the confidence interval of the indirect effect of cylinders on mpg

```
#Q1(c)
boot_mediation<-function(model1,model2,dataset) {
  boot_index<-sample(1:nrow(dataset), replace=TRUE)
  data_boot<-dataset[boot_index, ]
  regr1<-lm(model1,data_boot)
  regr2<-lm(model2,data_boot)
  return(regr1$coefficients[2] *regr2$coefficients[2])}
set.seed(42)
indirect<-replicate(2000,boot_mediation(wc,mw,cars_log))
quantile(indirect, probs=c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -0.7858081 -0.6565668
```

```
plot(density(indirect),lwd=2,col="blue", main='Bootstrapped Test of Indirect Effects between log.cylinders. on log.mp
g.')
abline(v=mean(indirect), lty=2, col="red", lwd=2)
abline(v=quantile(indirect, probs=c(0.025, 0.975)), lty=2, lwd=2, col="green")
```



Question 2

(a)Let's analyze the principal components of the four collinear variables

(i)Create a new data.frame of the four log-transformed variables with high multicollinearity

```
#Q2(a)(i)
engine <- with(cars_log, data.frame(log.mpg., log.cylinders., log.displacement., log.horsepower.))
engine<-na.omit(engine)
```

(ii)How much variance of the four variables is explained by their first principal component?

```
#Q2(a)(ii)
var<-eigen(cor(engine))$values
total_var<-sum(var)
var[1]/total_var
```

```
## [1] 0.8974062
```

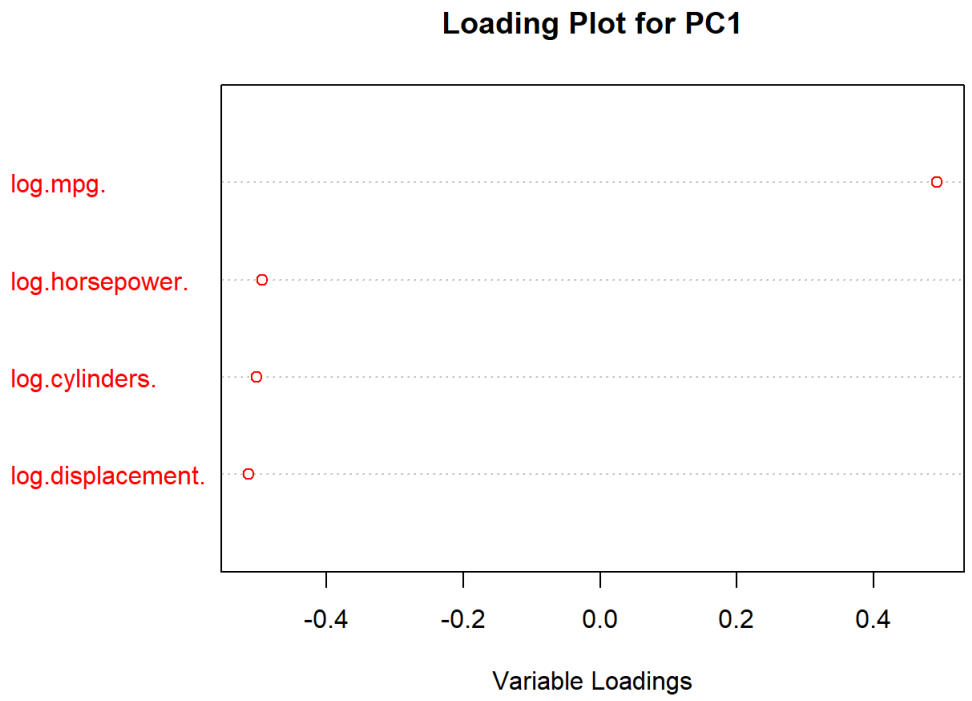
```
a<-prcomp(engine,scale. = T)
summary(a)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4
## Standard deviation     1.8946 0.46234 0.38683 0.21674
## Proportion of Variance 0.8974 0.05344 0.03741 0.01174
## Cumulative Proportion  0.8974 0.95085 0.98826 1.00000
```

89.72% of variance of the four variables is explained by their first principal component.

(iii)Looking at the values and valence (positive/negative) of the first principal component's eigenvector, what would you call the information captured by this component?

```
#Q2(a)(iii)
load1<-a$rotation[,1]
dotchart(load1[order(load1, decreasing=FALSE)] ,
         main="Loading Plot for PC1",
         xlab="Variable Loadings",
         col="red")
```



> The loading of log.mpg. in PC1 is positive, which means that it has the positive relationship with PC1. This tell us that PC1 can explain log.mpg very well as 89.72% of variance of the four variables is explained by PC1.

(b)Let's revisit our regression analysis on cars_log:

(i)Store the scores of the first principal component as a new column of cars_log

```
#Q2(b)(i)
PCs<-as.numeric(prcomp(engine, scale. = F)$x[,1])
cars_log$engine_PC1<-PCs
```

(ii)Regress mpg over the the column with PC1 scores

```
#Q2(b)(ii)
summary(lm(log.mpg.~engine_PC1+log.acceleration.+model_year+factor(origin),data=cars_log))
```

```
##
## Call:
## lm(formula = log.mpg. ~ engine_PC1 + log.acceleration. + model_year +
##     factor(origin), data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48398 -0.05446  0.00305  0.05678  0.44676
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.997440   0.156238  12.785  < 2e-16 ***
## engine_PC1         0.416275   0.012275  33.911  < 2e-16 ***
## log.acceleration. -0.230518   0.037283  -6.183  1.6e-09 ***
## model_year         0.022903   0.001657  13.823  < 2e-16 ***
## factor(origin)2   -0.036557   0.017952  -2.036   0.0424 *
## factor(origin)3   -0.020944   0.017694  -1.184   0.2373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1067 on 386 degrees of freedom
## Multiple R-squared:  0.9028, Adjusted R-squared:  0.9016
## F-statistic: 717.3 on 5 and 386 DF,  p-value: < 2.2e-16
```

(iii)Try running the regression again over the same independent variables, but this time with everything standardized. How important is this new column relative to other columns?

```
#Q2(b)(iii)
cars_log2<-as.data.frame(scale(cars_log[,1:7]))
PCs2<-as.numeric(prcomp(engine, scale. = T)$x[,1])
cars_log2$origin<-cars_log$origin
cars_log2$engine_PC1_std<-PCs2
summary(lm(log.mpg.~engine_PC1_std+log.acceleration.+model_year+factor(origin),data=cars_log2))
```

```
##
## Call:
## lm(formula = log.mpg. ~ engine_PC1_std + log.acceleration. +
##     model_year + factor(origin), data = cars_log2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31122 -0.15847 -0.00697  0.15401  1.12826
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.01581    0.02101   0.753    0.452
## engine_PC1_std     0.49080    0.01295  37.904  < 2e-16 ***
## log.acceleration. -0.13904    0.01831  -7.592  2.4e-13 ***
## model_year         0.21962    0.01669  13.161  < 2e-16 ***
## factor(origin)2   -0.07270    0.04762  -1.526    0.128
## factor(origin)3   -0.01589    0.04664  -0.341    0.734
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.288 on 386 degrees of freedom
## Multiple R-squared:  0.9181, Adjusted R-squared:  0.9171
## F-statistic: 865.6 on 5 and 386 DF,  p-value: < 2.2e-16
```

> The engine_PC1_std, log.acceleration., model_year are still significant effect on mpg, while the origins are not.

Question 3

(a)How much variance did each extracted factor explain?

```
#Q3(a)
library(readxl)
data<-read_excel("C:/Users/eva/Desktop/作業 上課資料(清大)/大四下/BACS/HW14 BACS/security_questions.xlsx","data")
data<-as.data.frame(data)
data_pca<-prcomp(data, scale. = T)
summary(data_pca)
```

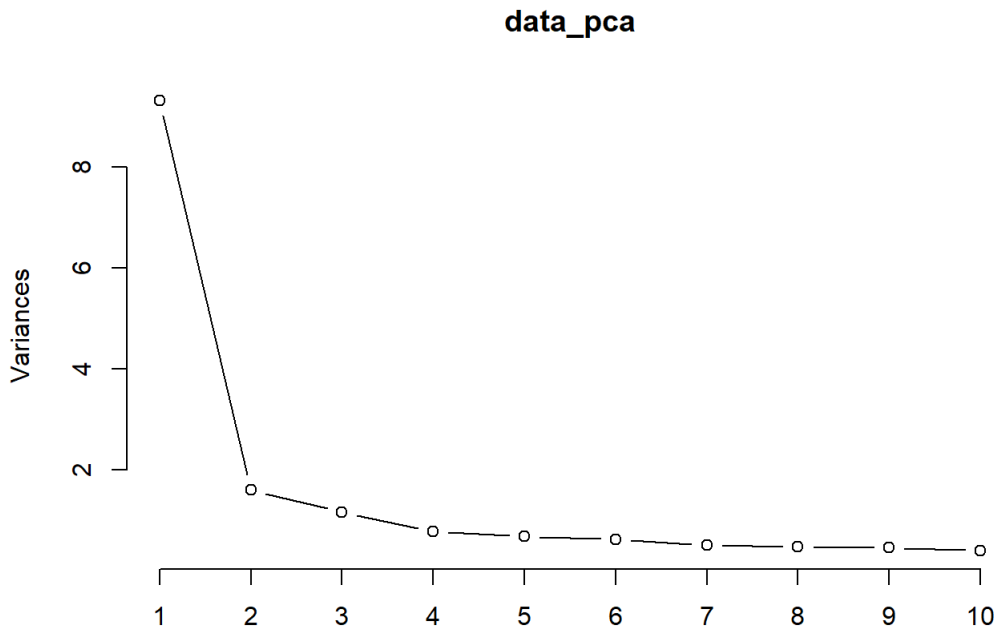
```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.0514 1.26346 1.07217 0.87291 0.82167 0.78209 0.70921
## Proportion of Variance 0.5173 0.08869 0.06386 0.04233 0.03751 0.03398 0.02794
## Cumulative Proportion  0.5173 0.60596 0.66982 0.71216 0.74966 0.78365 0.81159
##                           PC8     PC9   PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.68431 0.67229 0.6206 0.59572 0.54891 0.54063 0.51200
## Proportion of Variance 0.02602 0.02511 0.0214 0.01972 0.01674 0.01624 0.01456
## Cumulative Proportion  0.83760 0.86271 0.8841 0.90383 0.92057 0.93681 0.95137
##                           PC15   PC16   PC17   PC18
## Standard deviation     0.48433 0.4801 0.4569 0.4489
## Proportion of Variance 0.01303 0.0128 0.0116 0.0112
## Cumulative Proportion  0.96440 0.9772 0.9888 1.0000
```

(b)How many dimensions would you retain, according to the criteria we discussed?

```
#Q3(b)
eigen(cor(data))$values
```

```
##  [1] 9.3109533 1.5963320 1.1495582 0.7619759 0.6751412 0.6116636 0.5029855
##  [8] 0.4682788 0.4519711 0.3851964 0.3548816 0.3013071 0.2922773 0.2621437
## [15] 0.2345788 0.2304642 0.2087471 0.2015441
```

```
screeplot(data_pca, type="lines")
```



- Eigenvalues ≥ 1: retain 3 variables.
- Screeplot: retain 2 variables.

(c)(ungraded) Can you interpret what any of the principal components mean?

```
#Q3(c)
library(factoextra)
```

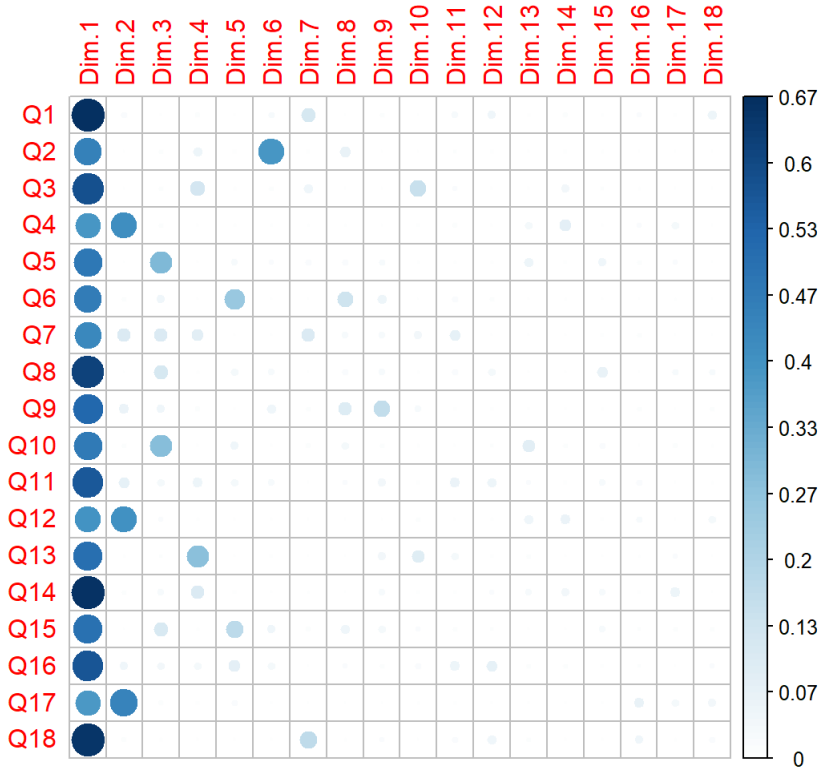
```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

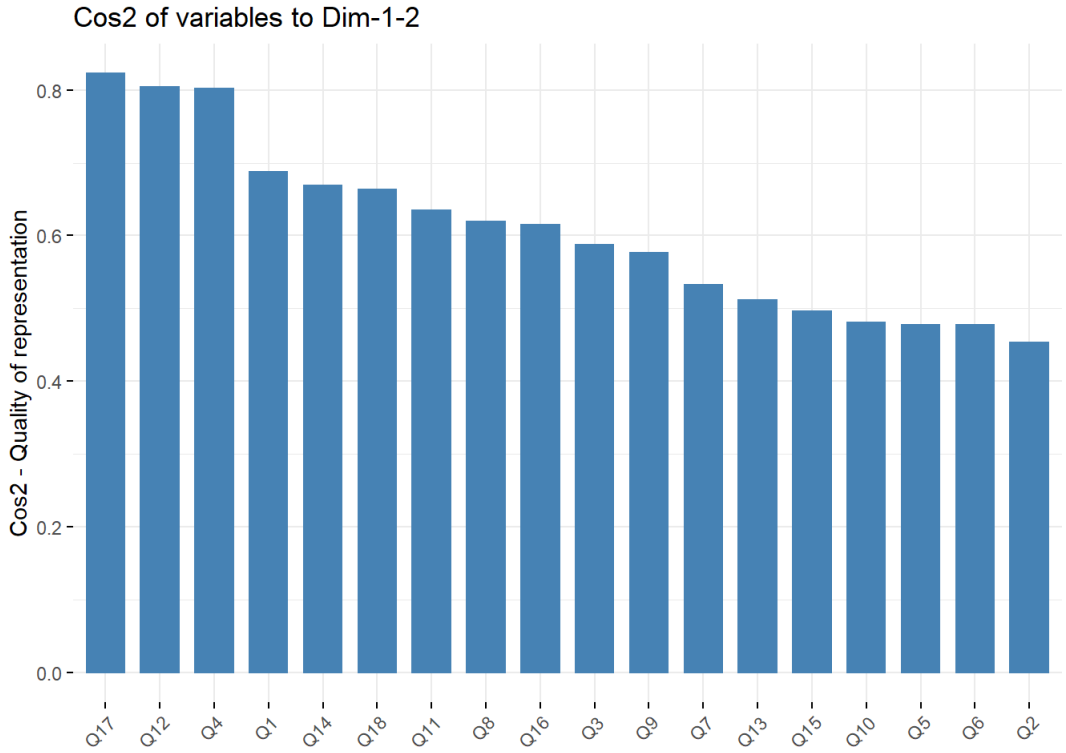
```
var <- get_pca_var(data_pca)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

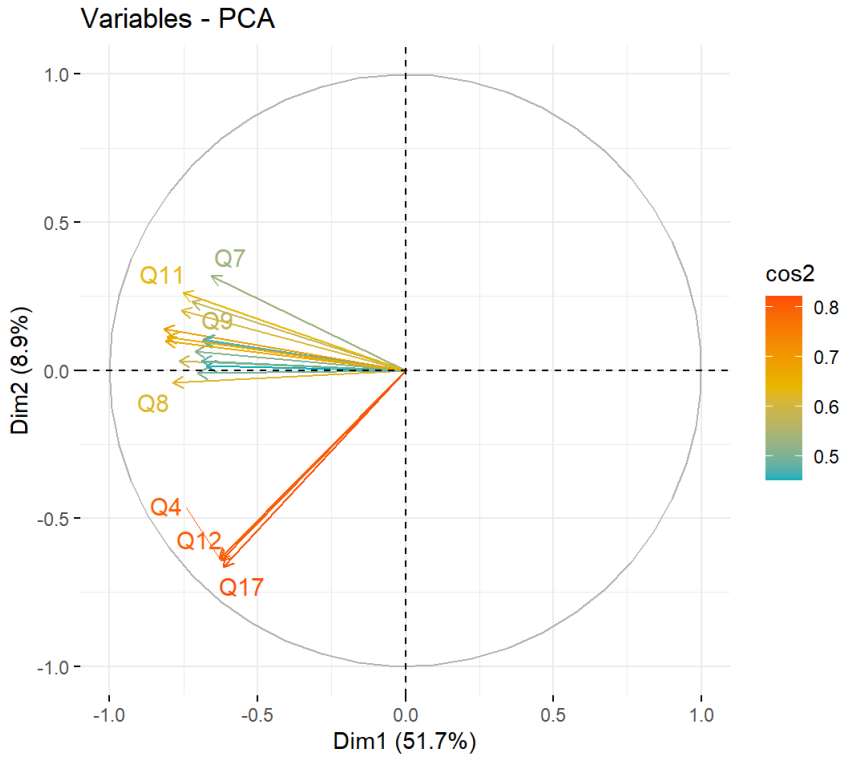
```
corrplot(var$cos2, is.corr=FALSE)
```




```
fviz_cos2(data_pca, choice = "var", axes = 1:2)
```




```
fviz_pca_var(data_pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping
             )
```



- A high cos2 value means that the variable has a large contribution to the principal component, which is represented in the example data by the correlation curve plotted near the edge of the circle.
- A low cos2 value means that the variable is not well represented by the principal component, which in the case of the correlation curve is approximately near the center of the circle.
- The cos2 value is to measure the usefulness of a variable. The higher it is, the more important the variable is in the principal component analysis.
- The variable correlation plot shows the relationship between the variables included in the group and the principal components, with positively correlated variables being close to each other and negatively correlated divisions being south of each other, while the length from the centroid to the variable represents the proportion of the variable in this dimension.

According to the correlation plot above, Q1, Q3, Q14, Q18 have greater effects on dim1. Q4, Q12, Q17 have greater effects on dim2. In dim-1-2, Q17 has largest cos2, and the following are Q12, Q4, Q1, Q14. And Q4, Q12, Q17 are highly positive correlated.

- (Dim2) Q4, Q12, Q17 are mainly about transaction, including transaction message and confirmation of transaction.
- (Dim1) Q1, Q3, Q14, Q18 are mainly about the security issues, including the accuracy of message, confidentiality of the transactions.