Exam-01

Pablo Vivas

2023-10-31

# Question 1

## Pre-work

First, we read the data and take a look at how does it look.

world = read\_csv("data/world-3.csv")  
  
world |>   
 sample\_n(10) |>   
 arrange(country) |>   
 gt() |>   
 fmt\_number(  
 columns = c(6, 10, 11, 12),  
 decimals = 2  
 )

| country | region | lifeexpf | lifeexpm | literacy | popincr | babymort | birthr | deathr | gdp | aidsr | bdratio | fertilty | literacym | literacyf |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Afghanistan | Pacific/Asia | 44 | 45 | 29 | 2.80 | 168.0 | 53 | 22.0 | 2.31 | 0.00 | 2.41 | 6.90 | 44 | 14 |
| Bulgaria | East Europe | 75 | 69 | 93 | -0.20 | 12.0 | 13 | 12.0 | 3.58 | 0.77 | 1.08 | 1.80 | NA | NA |
| Cameroon | Africa | 58 | 55 | 54 | 2.90 | 77.0 | 41 | 12.0 | 3.00 | 1.88 | 3.42 | 5.70 | 66 | 45 |
| Canada | OECD | 81 | 74 | 97 | 0.70 | 6.8 | 14 | 8.0 | 4.30 | 2.01 | 1.75 | 1.80 | NA | NA |
| Colombia | Latin America | 75 | 69 | 87 | 2.00 | 28.0 | 24 | 6.0 | 3.19 | 1.67 | 4.00 | 2.47 | 88 | 86 |
| Ecuador | Latin America | 73 | 67 | 88 | 2.01 | 39.0 | 26 | 6.0 | 3.04 | 1.28 | 4.33 | 3.08 | 90 | 86 |
| Italy | OECD | 81 | 74 | 97 | 0.21 | 7.6 | 11 | 10.0 | 4.24 | 2.07 | 1.10 | 1.30 | 98 | 96 |
| Jordan | Middle East | 74 | 70 | 80 | 3.30 | 34.0 | 39 | 5.0 | 3.06 | 0.93 | 7.80 | 5.64 | 89 | 70 |
| Paraguay | Latin America | 75 | 72 | 90 | 2.70 | 25.2 | 33 | 4.5 | 3.18 | 1.10 | 7.33 | 4.30 | 92 | 88 |
| Venezuela | Latin America | 76 | 70 | 88 | 2.16 | 28.0 | 26 | 5.0 | 3.45 | 1.75 | 5.20 | 3.05 | 90 | 87 |

**?(caption)**

Then, we make a description of all the numeric values.

world |>   
 select(where(is.numeric)) |>   
 skim() |>   
 select(-1) |>   
 gt() |>   
 fmt\_number(  
 columns = 3:10,  
 decimals = 2  
 ) |>   
 cols\_label(  
 skim\_variable = "variable",  
 n\_missing = "missing",  
 numeric.mean = "mean",  
 numeric.sd = "std",  
 numeric.p0 = "p0",  
 numeric.p25 = "p25",  
 numeric.p50 = "p50",  
 numeric.p75 = "p75",  
 numeric.p100 = "p100",  
 numeric.hist = "histogram"  
 ) |>   
 cols\_align(  
 columns = -1,  
 align = "center"  
 )

| variable | missing | complete\_rate | mean | std | p0 | p25 | p50 | p75 | p100 | histogram |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| lifeexpf | 0 | 1.00 | 70.16 | 10.57 | 43.00 | 67.00 | 74.00 | 78.00 | 82.00 | ▂▂▁▅▇ |
| lifeexpm | 0 | 1.00 | 64.92 | 9.27 | 41.00 | 61.00 | 67.00 | 72.00 | 76.00 | ▂▂▂▇▇ |
| literacy | 2 | 0.98 | 78.34 | 22.88 | 18.00 | 63.00 | 88.00 | 98.00 | 100.00 | ▁▂▂▂▇ |
| popincr | 0 | 1.00 | 1.68 | 1.20 | -0.30 | 0.52 | 1.80 | 2.68 | 5.24 | ▇▅▇▃▁ |
| babymort | 0 | 1.00 | 42.31 | 38.08 | 4.00 | 9.30 | 27.70 | 63.00 | 168.00 | ▇▃▁▂▁ |
| birthr | 0 | 1.00 | 25.92 | 12.36 | 10.00 | 14.00 | 25.00 | 35.00 | 53.00 | ▇▃▃▃▂ |
| deathr | 1 | 0.99 | 9.56 | 4.25 | 2.00 | 6.85 | 9.00 | 11.00 | 24.00 | ▅▇▅▁▁ |
| gdp | 0 | 1.00 | 3.42 | 0.62 | 2.09 | 3.00 | 3.48 | 3.87 | 4.37 | ▃▅▇▇▇ |
| aidsr | 3 | 0.97 | 1.38 | 0.71 | 0.00 | 0.77 | 1.38 | 1.86 | 3.18 | ▃▇▇▃▂ |
| bdratio | 1 | 0.99 | 3.20 | 2.12 | 0.92 | 1.54 | 2.67 | 4.18 | 14.00 | ▇▃▁▁▁ |
| fertilty | 2 | 0.98 | 3.56 | 1.90 | 1.30 | 1.88 | 3.05 | 5.00 | 8.19 | ▇▅▂▃▁ |
| literacym | 24 | 0.78 | 78.73 | 20.45 | 28.00 | 63.00 | 87.00 | 96.00 | 100.00 | ▁▁▃▂▇ |
| literacyf | 24 | 0.78 | 67.26 | 28.61 | 9.00 | 45.00 | 71.00 | 93.00 | 100.00 | ▂▂▂▃▇ |

**?(caption)**

We see from **?@tbl-desc-data** that some of the variables have missing data, specifically, the literacym variable has complete rate of 78%. Also, the standard deviation from each variable is in a different scale, ranging from 0.62 units in gdp to 38.08 units in babymort. This makes sense because some variables are in a log-scale.

|  |
| --- |
| Note |
| Given this scenario, we would prefer to perform pca with the correlation matrix. |

## Real work

### a)

We perform two principal component analysis using: a) the covariance matrix and b) the correlation matrix.

# get complete cases  
world\_complete = world |>   
 drop\_na()  
  
# for the covariance matrix  
pca\_cov = world\_complete |>   
 select(-c(1,2)) |>   
 prcomp()  
  
# for the correlation matrix  
pca\_cor = world\_complete |>   
 select(-c(1,2)) |>   
 prcomp(scale = TRUE)  
  
# matrix with loadings  
loadings = bind\_rows(  
 pca\_cov |>   
 tidy(matrix = "rotation") |>   
 mutate(matrix = factor("covariance")),  
 pca\_cor |>   
 tidy(matrix = "rotation") |>   
 mutate(matrix = factor("correlation"))  
)

We plot the differences for the first two pcs using each method.

Option 1:

plot\_01 = tibble(cor = pca\_cor$rotation[,1], cov = pca\_cov$rotation[,1])   
plot\_02 = tibble(cor = pca\_cor$rotation[,2], cov = pca\_cov$rotation[,2])   
  
p01 = plot\_01 |>   
 ggplot() +  
 geom\_point(aes(x = cor, y = cov), size = 3) +  
 geom\_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red", size = 1.3) +  
 xlim(c(-0.65, 0.65)) +  
 ylim(c(-0.65, 0.65)) +  
 theme\_bw() +  
 xlab("PC1 (cor)") +  
 ylab("PC1 (cov)")  
  
p02 = plot\_02 |>   
 ggplot() +  
 geom\_point(aes(x = cor, y = cov), size = 3) +  
 geom\_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red", size = 1.3) +  
 xlim(c(-0.65, 0.65)) +  
 ylim(c(-0.65, 0.65)) +  
 theme\_bw() +  
 xlab("PC2 (cor)") +  
 ylab("PC2 (cov)")  
  
p01 | p02

|  |
| --- |
| Figure 1: Representations of the loadings |

Option 2:

loadings |>   
 filter(PC == c(1,2)) |>   
 ggplot() +  
 geom\_density(aes(x = value, fill = matrix, color = matrix), alpha = 0.6) +  
 facet\_wrap(vars(PC), nrow = 1) +  
 theme\_bw() +  
 theme(legend.position = "bottom") +  
 scale\_color\_brewer(palette = "Pastel1") +  
 scale\_fill\_brewer(palette = "Pastel1")

|  |
| --- |
| Figure 2: Representations of the loadings |

Option 3:

loadings |>   
 filter(PC == c(1,2)) |>   
 ggplot() +  
 geom\_boxplot(aes(x = value, y = matrix, fill = matrix, color = matrix), alpha = 0.6) +  
 facet\_wrap(vars(PC), nrow = 1) +  
 theme\_bw() +  
 theme(legend.position = "none") +  
 scale\_color\_brewer(palette = "Accent") +  
 scale\_fill\_brewer(palette = "Accent")

|  |
| --- |
| Figure 3: Representations of the loadings |

Option 4:

p03 = pca\_cov %>%  
 tidy(matrix = "eigenvalues") %>%  
 ggplot(aes(PC, percent)) +  
 geom\_col(fill = "#56B4E9", alpha = 0.8) +  
 xlab("PCA (cov)") +  
 scale\_x\_continuous(breaks = 1:9) +  
 scale\_y\_continuous(  
 labels = scales::percent\_format(),  
 expand = expansion(mult = c(0, 0.01))  
 ) +  
 theme\_minimal\_hgrid(12)  
  
p04 = pca\_cor %>%  
 tidy(matrix = "eigenvalues") %>%  
 ggplot(aes(PC, percent)) +  
 xlab("PCA (cor)") +  
 geom\_col(fill = "#56B4E9", alpha = 0.8) +  
 scale\_x\_continuous(breaks = 1:9) +  
 scale\_y\_continuous(  
 labels = scales::percent\_format(),  
 expand = expansion(mult = c(0, 0.01))  
 ) +  
 theme\_minimal\_hgrid(12)  
  
p03 | p04

|  |
| --- |
| Figure 4: Representations of the loadings |

### b)

### c)

### d)

world\_complete = pca\_cor |>  
 augment(world\_complete)

# Question 2

# Question 3