Exam-01

Pablo Vivas

2023-11-04

# Question 1

## Descriptive analysis

First, we read the data and take a look at some observations.

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 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Bangladesh | Pacific/Asia | 53 | 53 | 35 | 2.40 | 106.0 | 35 | 11 | 2.31 | 0.24 | 3.18 | 4.70 | 47 | 22 |
| Cambodia | Pacific/Asia | 52 | 50 | 35 | 2.90 | 112.0 | 45 | 16 | 2.41 | 0.00 | 2.81 | 5.81 | 48 | 22 |
| China | Pacific/Asia | 69 | 67 | 78 | 1.10 | 52.0 | 21 | 7 | 2.58 | 0.32 | 3.00 | 1.84 | 87 | 68 |
| Ecuador | Latin America | 73 | 67 | 88 | 2.01 | 39.0 | 26 | 6 | 3.04 | 1.28 | 4.33 | 3.08 | 90 | 86 |
| Georgia | East Europe | 76 | 69 | 99 | 0.80 | 23.0 | 16 | 9 | 3.65 | 0.52 | 1.78 | 2.18 | 100 | 100 |
| Lithuania | East Europe | 77 | 68 | 99 | 0.30 | 17.0 | 15 | 10 | 3.83 | 0.67 | 1.50 | 2.00 | 99 | 98 |
| Malaysia | Pacific/Asia | 72 | 66 | 78 | 2.30 | 25.6 | 29 | 5 | 3.48 | 0.89 | 5.80 | 3.51 | 86 | 70 |
| Singapore | Pacific/Asia | 79 | 73 | 88 | 1.20 | 5.7 | 16 | 6 | 4.18 | 1.21 | 2.67 | 1.88 | 93 | 84 |
| Vietnam | Pacific/Asia | 68 | 63 | 88 | 1.78 | 46.0 | 27 | 8 | 2.36 | 0.68 | 3.38 | 3.33 | 93 | 83 |
| Zambia | Africa | 45 | 44 | 73 | 2.80 | 85.0 | 46 | 18 | 2.76 | 3.18 | 2.56 | 6.68 | 81 | 65 |

**?(caption)**

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

world |>   
 select(where(is.numeric)) |>   
 skim() |>   
 select(-1) |>   
 arrange(skim\_variable) |>   
 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 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| aidsr | 3 | 0.97 | 1.38 | 0.71 | 0.00 | 0.77 | 1.38 | 1.86 | 3.18 | ▃▇▇▃▂ |
| babymort | 0 | 1.00 | 42.31 | 38.08 | 4.00 | 9.30 | 27.70 | 63.00 | 168.00 | ▇▃▁▂▁ |
| bdratio | 1 | 0.99 | 3.20 | 2.12 | 0.92 | 1.54 | 2.67 | 4.18 | 14.00 | ▇▃▁▁▁ |
| 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 | ▅▇▅▁▁ |
| fertilty | 2 | 0.98 | 3.56 | 1.90 | 1.30 | 1.88 | 3.05 | 5.00 | 8.19 | ▇▅▂▃▁ |
| gdp | 0 | 1.00 | 3.42 | 0.62 | 2.09 | 3.00 | 3.48 | 3.87 | 4.37 | ▃▅▇▇▇ |
| 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 | ▁▂▂▂▇ |
| literacyf | 24 | 0.78 | 67.26 | 28.61 | 9.00 | 45.00 | 71.00 | 93.00 | 100.00 | ▂▂▂▃▇ |
| literacym | 24 | 0.78 | 78.73 | 20.45 | 28.00 | 63.00 | 87.00 | 96.00 | 100.00 | ▁▁▃▂▇ |
| popincr | 0 | 1.00 | 1.68 | 1.20 | -0.30 | 0.52 | 1.80 | 2.68 | 5.24 | ▇▅▇▃▁ |

**?(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. |

### 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 pca’s using each method.

data\_p01 = tibble(var = colnames(world\_complete)[-c(1,2)], pca\_1 = pca\_cov$rotation[,1], pca\_2 = pca\_cov$rotation[,2])   
  
data\_p02 = tibble(var = colnames(world\_complete)[-c(1,2)], pca\_1 = pca\_cor$rotation[,1], pca\_2 = pca\_cor$rotation[,2])   
  
p01 = data\_p01 |>   
 ggplot(aes(x = pca\_1, y = pca\_2, label= var)) +  
 geom\_point(size = 4, color = "red") +  
 geom\_label\_repel(size = 4) +  
 xlim(c(-0.65, 0.50)) +  
 ylim(c(-0.65, 0.65)) +  
 xlab("PC1 (cov)") +  
 ylab("PC2 (cov)") +  
 theme\_classic()  
  
p02 = data\_p02 |>   
 ggplot(aes(x = pca\_1, y = pca\_2, label= var)) +  
 geom\_point(size = 4, color = "red") +  
 geom\_label\_repel(size = 4) +  
 xlim(c(-0.65, 0.50)) +  
 ylim(c(-0.65, 0.65)) +  
 xlab("PC1 (cor)") +  
 ylab("PC2 (cor)") +  
 theme\_classic()  
  
p01 | p02

|  |
| --- |
| Figure 1: Plot of principal component loadings for the first two pcs using the covariance matrix (left) and the correlation matrix (right) |

Based on the results presented on [Figure 1](#fig-plot-comparison) we can conclude that the loadings across the two solutions differ (although the plot only presented the first two components for each solution, the same behavior, or a lesser extent of it, is observed on the remainder of the components). The right plot shows that the variables related to literacy have larger values on the first and second principal component than the other variables (except for babymort in the second component). This behavior matches the difference in standard deviation we see in **?@tbl-desc-data**. On the other hand, the left plot shows a more distributed pattern of the loadings, where no variable stands out from the rest. The difference between the loadings across the two solutions can be explain by the scale-dependency of the technique. Given this difference, I think it is more appropriate to analyze this dataset with the correlation matrix.

### b)

To investigate the number of necessary principal components to adequately describe the variability among the variables, several approaches will be taken.

1. Scree-plot & Kaiser rule

pca\_cor |>   
 tidy(matrix = "eigenvalues") |>   
 mutate(lambda = std.dev \* std.dev) |>   
 ggplot(aes(PC, lambda)) +  
 geom\_col(fill = "#56B4E9", alpha = 0.8) +  
 geom\_abline(slope = 0, intercept = 1, linetype = "dashed", color = "red") +  
 xlab("PCA") +  
 scale\_x\_continuous(breaks = 1:13) +  
 scale\_y\_continuous(  
 expand = expansion(mult = c(0, 0.01))  
 ) +  
 theme\_minimal\_hgrid(12)

|  |
| --- |
| Figure 2: Scree plot |

The scree plot suggests that we need to extract two principal components, because at pca number 2 we notice the “elbow” in the amount of variance explained. After this component, the remainder pcas account for a negligible amount of variance. Plus, the Kaiser rule also suggest that we need to extract two principal components because only the first two have an eigenvalue greater than one.

1. Proportion of explained variance.

pca\_cor |>   
 tidy(matrix = "eigenvalues") |>  
 mutate(per = round(percent \* 100, 2),  
 per.cum = round(cumulative \* 100, 2),  
 .keep = "unused") |>   
 select(-std.dev) |>   
 gt() |>   
 cols\_label(  
 per = "Percentage",  
 per.cum = "Cummulative"  
 ) |>   
 cols\_align(  
 columns = -1,  
 align = "center"  
 )

| PC | Percentage | Cummulative |
| --- | --- | --- |
| 1 | 64.76 | 64.76 |
| 2 | 18.00 | 82.75 |
| 3 | 7.30 | 90.05 |
| 4 | 4.31 | 94.36 |
| 5 | 2.55 | 96.92 |
| 6 | 0.99 | 97.91 |
| 7 | 0.64 | 98.55 |
| 8 | 0.43 | 98.98 |
| 9 | 0.40 | 99.37 |
| 10 | 0.29 | 99.66 |
| 11 | 0.15 | 99.81 |
| 12 | 0.13 | 99.94 |
| 13 | 0.06 | 100.00 |

**?(caption)**

Based on **?@tbl-percentage**, the analysis indicates that two principal components explain over a 80% of the variance. Even thought is subjective, I believe that this percentage is good enough for this context and that two components explaining over a 80% in this analysis with 13 variables is acceptable.

|  |
| --- |
| Note |
| The three methods suggest that 2 principal components appear to be necessary to describe adequately the variability in these variables. The rest of the analysis will be done using two principal components. |

### c)

PCA 1 =

Countries that would receive high scores in this pc will be those. Countries that would receive low scores in this pc will be those

PCA 2 =

Countries that would receive high scores in this pc will be those. Countries that would receive low scores in this pc will be those

### d)

world\_complete = pca\_cor |>  
 augment(world\_complete)  
  
world\_complete |>   
 ggplot() +  
 geom\_point(aes(x = .fittedPC1, y = .fittedPC2, color = region), size = 3) +  
 xlab("PC1") +  
 ylab("PC2") +  
 theme\_classic()

|  |
| --- |
| Figure 3: Plot of something |

### e)

pca\_cov$rotation[,1]

lifeexpf lifeexpm literacy popincr babymort birthr   
 0.174578148 0.144738162 0.394659468 -0.011644817 -0.647106910 -0.183190702   
 deathr gdp aidsr bdratio fertilty literacym   
-0.053528089 0.007484552 -0.001743288 -0.001454908 -0.028573196 0.331119528   
 literacyf   
 0.476437248

# Question 2

# Question 3