Exam-01

Pablo Vivas

2023-11-07

# 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 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Afghanistan | Pacific/Asia | 44 | 45 | 29 | 2.80 | 168.0 | 53 | 22.0 | 2.31 | 0.00 | 2.41 | 6.90 | 44 | 14 |
| Croatia | East Europe | 77 | 70 | 97 | -0.10 | 8.7 | 11 | 11.0 | 3.74 | 1.03 | 1.00 | 1.65 | NA | NA |
| Czech Rep. | East Europe | 77 | 69 | NA | 0.21 | 9.3 | 13 | 11.1 | 3.86 | 0.86 | 1.17 | 1.84 | NA | NA |
| Ethiopia | Africa | 54 | 51 | 24 | 3.10 | 110.0 | 45 | 14.0 | 2.09 | 1.88 | 3.21 | 6.81 | 32 | 16 |
| Iran | Middle East | 67 | 65 | 54 | 3.46 | 60.0 | 42 | 8.0 | 3.18 | 0.68 | 5.25 | 6.33 | 64 | 43 |
| Ireland | OECD | 78 | 73 | 98 | 0.30 | 7.4 | 14 | 9.0 | 4.09 | 1.61 | 1.56 | 1.99 | NA | NA |
| Netherlands | OECD | 81 | 75 | 99 | 0.58 | 6.3 | 13 | 9.0 | 4.24 | 1.82 | 1.44 | 1.58 | NA | NA |
| Somalia | Africa | 55 | 54 | 24 | 3.20 | 126.0 | 46 | 13.0 | 3.33 | 0.67 | 3.54 | 7.25 | 36 | 14 |
| Uruguay | Latin America | 77 | 71 | 96 | 0.80 | 17.0 | 17 | 10.0 | 3.50 | 1.71 | 1.70 | 2.44 | 97 | 96 |
| 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) |>   
 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 and the literacyf variables have 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 of differences between scales and variability, we would prefer to perform the principal component analysis with the correlation matrix instead of the covariance 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)

We plot the differences for the two analysis using each method. The plots will contain the first two principal components.

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 and Kaiser rule (dotted red line) |

The scree plot suggests that we need to extract two principal components, because at pca number 3 we notice the “elbow” in the amount of variance explained. After the second component, the remainder principal components 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\_width(  
 per ~ px(120),  
 per.cum ~ px(120)  
 ) |>  
 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. Also, two components explaining over a 80% in this analysis with 13 variables is acceptable/awesome 😁.

|  |
| --- |
| Note |
| The three methods (scree-plot, Kaiser rule and proportion of variance explained) 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)

In order to make interpretations of the two principal components, we will take a look at the loadings:

data\_p02 |>   
 gt() |>   
 cols\_width(  
 everything() ~ px(150)  
 ) |>   
 cols\_align(  
 columns = -1,  
 align = "center"  
 ) |>   
 fmt\_number(  
 columns = 2:3,  
 decimals = 2  
 ) |>   
 cols\_label(  
 var = "Variable",  
 pca\_1 = "PC1",  
 pca\_2 = "PC2"  
 )

| Variable | PC1 | PC2 |
| --- | --- | --- |
| lifeexpf | 0.33 | -0.15 |
| lifeexpm | 0.31 | -0.22 |
| literacy | 0.32 | 0.05 |
| popincr | -0.23 | -0.46 |
| babymort | -0.33 | 0.11 |
| birthr | -0.32 | -0.16 |
| deathr | -0.23 | 0.44 |
| gdp | 0.27 | -0.05 |
| aidsr | -0.07 | 0.24 |
| bdratio | -0.03 | -0.62 |
| fertilty | -0.32 | -0.15 |
| literacym | 0.31 | 0.07 |
| literacyf | 0.32 | 0.07 |

**?(caption)**

Based on **?@tbl-loadings-cor** and its graphical representation in [Figure 1](#fig-plot-comparison) (right plot) we can interpret each principal component as follows:

* *Principal Component 1:* This component seems to be associated with general or overall development indicators. High positive scores on PC1 correlate with high GDP, high literacy rates (for both males and females), and higher life expectancy for both males and females. This suggests that countries with higher scores on PC1 are likely to be more developed, with stronger economies and better health and education systems. On the other hand, countries with low scores on PC1 exhibit higher rates of infant mortality (babymort), higher AIDS rates (aidsr), and higher birth rates (birthr). These are typical of less developed countries where healthcare and education systems are weaker and economic activity is lower.
* *Principal Component 2:* This component captures more demographic variables and aspects of population dynamics. Notably, countries with high scores on PC2 have higher death rates (deathr) and possibly lower population increases (popincr), indicating countries that may be experiencing either low birth rates, high mortality rates, or a combination of both. On the other end, countries with low scores on PC2 have higher fertility, birth rates (birthr), and a higher birth to death ratio (bdratio), suggesting these are countries with younger populations and potentially higher population growth due to these factors.

### 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") +  
 scale\_color\_brewer(palette = "Dark2") +  
 theme\_classic()

|  |
| --- |
| Figure 3: Plot of principal component by region |

Based on [Figure 3](#fig-pca-01), there does appear to be an association between the component scores and the region a country belongs to. For instance:

* OECD countries are generally clustered towards higher values on PC1. This suggests that OECD countries, which are typically more economically developed, have higher GDP, literacy rates, and life expectancy, characteristics that are positively correlated with PC1.
* African countries are spread across both PC1 and PC2, but they are generally clustered towards lower PC1 values, indicating lower GDP, literacy rates, and life expectancy. The spread across PC2 might indicate variability in demographic patterns within the region, such as differences in birth and death rates.
* Latin American, East European, Middle Eastern, and Pacific/Asian countries appear to be intermediate on PC1, which would suggest these regions have varying levels of development. They’re not as high as OECD countries but not as low as many African countries.
* Latin American countries show a slight trend towards lower values on PC2, which could imply higher birth rates and fertility rates in these countries.
* East European countries seem to be more variable along PC2, which might reflect a diversity in demographic transitions, possibly due to the varying economic and historical contexts within this region.
* Middle Eastern and Pacific/Asian countries are scattered across both components, indicating a wide variety of economic, health, and demographic profiles within these regions.

### e)

To show that the weight vectors for the first two components are orthogonal, we first need to extract the weight vectors or loadings for the perform analysis:

loadings = pca\_cov$rotation

Then we check the dot product for the first two loadings vectors

round(sum(loadings[, 1] \* loadings[, 2]), 3)

[1] 0

As expected, this value is , showing that the weight or loadings for the first two components are orthogonal.

# Question 2

We read the data.

wais = matrix(c(1.00, .37, .34, .40, .27, .59, .09, .25, .27, .22, .26,  
 .37, 1.00, .27, .25, .38, .46, .10, .26, .29, .22, .24,  
 .34, .27, 1.00, .36, .28, .33, .18, .32, .38, .29, .30,  
 .40, .25, .36, 1.00, .22, .35, .08, .31, .26, .25, .20,  
 .27, .38, .28, .22, 1.00, .29, .16, .14, .18, .15, .22,  
 .59, .46, .33, .35, .29, 1.00, .08, .27, .24, .28, .26,  
 .09, .10, .18, .08, .16, .08, 1.00, .19, .13, .22, .17,  
 .25, .26, .32, .31, .14, .27, .19, 1.00, .36, .36, .40,  
 .27, .29, .38, .26, .18, .24, .13, .36, 1.00, .30, .60,   
 .22, .22, .29, .25, .15, .28, .22, .36, .30, 1.00, .25,  
 .26, .24, .30, .20, .22, .26, .17, .40, .60, .25, 1.00),   
ncol=11, byrow=T)  
  
rownames(wais) = colnames(wais) = c("Information", "Comprehension",   
"Arithmetic", "Similarities", "Digit Span", "Vocabulary", "Digit Symbol",   
"Picture Completion", "Block Design", "Picture Arrangement", "Object   
Assembly")

### a)

We perform a pca with the correlation matrix

#wais\_pc = prcomp(covmat = wais)

# wais\_pc |>   
# 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)

# Question 3