Exam-02

Pablo Vivas

2023-12-14

Table of contents

# 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 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Barbados | Latin America | 78 | 73 | 99 | 0.21 | 20.3 | 16 | 8.4 | 3.84 | 2.68 | 1.90 | 1.78 | 99 | 99 |
| Bulgaria | East Europe | 75 | 69 | 93 | -0.20 | 12.0 | 13 | 12.0 | 3.58 | 0.77 | 1.08 | 1.80 | NA | NA |
| Costa Rica | Latin America | 79 | 76 | 93 | 2.30 | 11.0 | 26 | 4.0 | 3.31 | 1.78 | 6.50 | 3.10 | 93 | 93 |
| Cuba | Latin America | 78 | 74 | 94 | 0.95 | 10.2 | 17 | 7.0 | 3.14 | 1.17 | 2.43 | 1.90 | 95 | 93 |
| Greece | OECD | 80 | 75 | 93 | 0.84 | 8.2 | 10 | 10.0 | 3.91 | 1.55 | 1.00 | 1.50 | 98 | 89 |
| Guatemala | Latin America | 67 | 62 | 55 | 2.58 | 57.0 | 35 | 8.0 | 3.13 | 1.37 | 4.38 | 4.76 | 63 | 47 |
| Indonesia | Pacific/Asia | 65 | 61 | 77 | 1.60 | 68.0 | 24 | 9.0 | 2.83 | 0.48 | 2.67 | 2.80 | 84 | 68 |
| Ireland | OECD | 78 | 73 | 98 | 0.30 | 7.4 | 14 | 9.0 | 4.09 | 1.61 | 1.56 | 1.99 | NA | NA |
| South Africa | Africa | 68 | 62 | 76 | 2.60 | 47.1 | 34 | 8.0 | 3.50 | 1.51 | 4.25 | 4.37 | NA | NA |
| Switzerland | OECD | 82 | 75 | 99 | 0.70 | 6.2 | 12 | 9.0 | 4.35 | 2.21 | 1.33 | 1.60 | NA | NA |

**?(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 |>   
 mutate(Region = region) |>   
 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 and East Europe countries are generally clustered towards higher values on PC1. This suggests that these 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, 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 or East Europe 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.
* 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 and evaluate and identify an appropriate number of principal components with the scree-plot, the Kaiser rule and the proportion of variance explained.

wais\_pc = princomp(covmat = wais)  
  
screeplot(wais\_pc); abline(h = 1, col = "red")

|  |
| --- |
| Figure 4: Scree plot |

The scree-plot and the Kaiser rule suggests that 2 principal components would be appropriate to describe the variance among the variables. Now we compute the proportion of variance explained by each component.

lambda = wais\_pc$sdev^2  
  
v = NULL  
for (i in 1:11) v[i] = lambda[i]/11  
  
tibble(  
 PC = c(1:11),  
 per = round(v\*100, 2),  
 cum = round(cumsum(lambda)/11 \*100, 2)   
 ) |>   
 gt() |>  
 cols\_width(  
 per ~ px(120),  
 cum ~ px(120)  
 ) |>  
 cols\_label(  
 per = "Percentage",  
 cum = "Cummulative"  
 ) |>   
 cols\_align(  
 columns = -1,  
 align = "center"  
 )

| PC | Percentage | Cummulative |
| --- | --- | --- |
| 1 | 34.46 | 34.46 |
| 2 | 11.61 | 46.07 |
| 3 | 8.99 | 55.06 |
| 4 | 8.40 | 63.46 |
| 5 | 7.00 | 70.46 |
| 6 | 6.49 | 76.96 |
| 7 | 5.86 | 82.82 |
| 8 | 5.36 | 88.18 |
| 9 | 5.01 | 93.19 |
| 10 | 3.63 | 96.82 |
| 11 | 3.18 | 100.00 |

**?(caption)**

Based on **?@tbl-pc-results**, we see that the first component explained around 34% of the variance, the second and the third component around 11% and 9%, respectively. This information doesn’t provide much evidence toward the number of principal components to extract. Hence, we rely on the scree-plot and the Kaiser rule. Therefore, 2 principal components would be appropriate to describe the variance among the variables. And together, this two components explained about 46% of the variance.

|  |
| --- |
| Note |
| The rest of the analysis will be done using two principal components. |

### b)

A = loadings(wais\_pc)  
  
loadings=NULL  
for (i in 1:length(lambda)){  
 loadings = cbind(loadings, A[,i]\* sqrt(lambda[i]))  
}  
  
tibble(  
 Variable = row.names(loadings),  
 PC1 = loadings[,1],  
 PC2 = loadings[,2]  
) |>   
 gt()|>   
 cols\_width(  
 everything() ~ px(150)  
 ) |>   
 cols\_align(  
 columns = -1,  
 align = "center"  
 ) |>   
 fmt\_number(  
 columns = 2:3,  
 decimals = 2  
 )

| Variable | PC1 | PC2 |
| --- | --- | --- |
| Information | 0.66 | 0.42 |
| Comprehension | 0.61 | 0.33 |
| Arithmetic | 0.64 | -0.05 |
| Similarities | 0.58 | 0.19 |
| Digit Span | 0.49 | 0.29 |
| Vocabulary | 0.67 | 0.44 |
| Digit Symbol | 0.30 | -0.34 |
| Picture Completion | 0.60 | -0.34 |
| Block Design | 0.64 | -0.42 |
| Picture Arrangement | 0.54 | -0.26 |
| Object Assembly | 0.62 | -0.45 |

**?(caption)**

Based on **?@tbl-loadings-cor-02** we can interpret each principal component as follows:

* **Principal Component 1:** seems to capture general intelligence (or as our friend Spearman would called it the “g” factor) since all the subtests have positive loadings and most of them are strong. This suggests that PC1 might be measuring a general cognitive ability that affects performance across all the subtests.
* **Principal Component 2:** this component distinguishes between subtests that involve verbal comprehension and those that require other skills such as processing speed, nonverbal reasoning, or spatial processing. Subtests like “Information”, “Comprehension”, and “Vocabulary” have positive loadings, which may represent a ‘verbal’ dimension. In contrast, tests such as “Block Design”, “Picture Completion”, and “Object Assembly” have negative loadings, possibly indicating a ‘performance’ dimension or nonverbal skills.

### c)

fa\_01 = fa(wais, nfactor = 1, n.obs = 230, fm = "ml", rotate = "none")  
fa\_02 = fa(wais, nfactor = 2, n.obs = 230, fm = "ml", rotate = "none")  
fa\_03 = fa(wais, nfactor = 3, n.obs = 230, fm = "ml", rotate = "none")  
fa\_04 = fa(wais, nfactor = 4, n.obs = 230, fm = "ml", rotate = "none")  
  
tibble(  
 Model = str\_c(1:4,"-factor"),  
 Pval = c(fa\_01$PVAL, fa\_02$PVAL, fa\_03$PVAL, fa\_04$PVAL),  
 TLI = c(fa\_01$TLI, fa\_02$TLI, fa\_03$TLI, 1),  
 RMSEA = c(fa\_01$RMSEA[1], fa\_02$RMSEA[1], fa\_03$RMSEA[1], fa\_04$RMSEA[1])  
) |>   
 gt() |>   
 fmt\_number(  
 columns = 2:4,  
 decimals = 3  
 ) |>   
 cols\_width(  
 everything() ~ px(80)  
 )

| Model | Pval | TLI | RMSEA |
| --- | --- | --- | --- |
| 1-factor | 0.000 | 0.775 | 0.098 |
| 2-factor | 0.022 | 0.944 | 0.049 |
| 3-factor | 0.170 | 0.973 | 0.034 |
| 4-factor | 0.484 | 1.000 | 0.000 |

**?(caption)**

Based on **?@tbl-factor-results** we can conclude the following:

* 1-factor model: This model does not appear to fit the data well. The P-value is very low, it could imply that the model significantly differs from the observed data. This is supported by a TLI that is below the acceptable threshold of 0.90 and an RMSEA that is considered high, indicating a poor fit.
* 2-factor model: This model shows significant improvement over the 1-factor model. The P-value is higher, although it’s still below 0.05. The TLI is much improved and is above the threshold of 0.90, indicating a good fit. The RMSEA is within the range of what is considered a reasonable error of approximation, suggesting this model fits the data well.
* 3-factor model: This model continues the trend of improved fit. The P-value is well above 0.05, suggesting that the model fits the data closely enough that any differences from the observed data are not statistically significant. The TLI is very close to 1, and the RMSEA is well below 0.05, which indicates a close fit.
* 4-factor model: This model has the best fit according to the indices presented. The P-value is high, which could suggest the model fits the data so well that there’s no difference between the model and the observed data, although it could also indicate overfitting. The TLI is perfect, and the RMSEA is 0, which might actually raise suspicions as it suggests a perfect model fit, which is very rare in practice.

Using a parsimonious approach, we can use the 2-factor model because it provides a balance between complexity and goodness of fit.

tibble(  
 Variable = row.names(loadings),  
 FA1 = fa\_02$loadings[, 1],  
 FA2 = fa\_02$loadings[, 2]  
) |>   
 gt()|>   
 cols\_width(  
 everything() ~ px(150)  
 ) |>   
 cols\_align(  
 columns = -1,  
 align = "center"  
 ) |>   
 fmt\_number(  
 columns = 2:3,  
 decimals = 2  
 )

| Variable | FA1 | FA2 |
| --- | --- | --- |
| Information | 0.64 | -0.35 |
| Comprehension | 0.55 | -0.18 |
| Arithmetic | 0.56 | 0.03 |
| Similarities | 0.50 | -0.13 |
| Digit Span | 0.41 | -0.12 |
| Vocabulary | 0.66 | -0.41 |
| Digit Symbol | 0.22 | 0.09 |
| Picture Completion | 0.52 | 0.16 |
| Block Design | 0.64 | 0.44 |
| Picture Arrangement | 0.45 | 0.06 |
| Object Assembly | 0.61 | 0.42 |

**?(caption)**

Based on **?@tbl-loadings-fa**, we can say the following:

* **Factor 1** seems to represent a broad cognitive ability that might be related to general intelligence, given that it is positively associated with a wide range of subtests, particularly those requiring verbal comprehension, problem-solving, and perceptual organization. This is consistent with the general intelligence (“g”) factor framework (once again, our friend Spearman).
* **Factor 2** appears to distinguish between subtests that require visual-spatial processing and manipulation (e.g., Block Design, Object Assembly) which have positive loadings, and subtests that are more verbally oriented (e.g., Vocabulary, Information) which have negative loadings. This suggests that Factor 2 could be capturing a contrast between visual-spatial abilities and verbal abilities.

The PCA results showed that all subtests had positive loadings on PC1, suggesting a single underlying dimension. On PC2, there was a clear distinction between subtests, with those requiring verbal comprehension (e.g., Information, Comprehension, Vocabulary) having positive loadings, and those requiring processing speed or non-verbal reasoning (e.g., Digit Symbol, Picture Completion, Block Design) having negative loadings.

In the factor analysis results:

Factor 1 seems to align with the PC1 from the PCA, where it represents a general intelligence factor given its positive association with a range of subtests.Factor 2 appears to align somewhat with PC2 from the PCA in terms of distinguishing between verbal and non-verbal abilities but with a different pattern of loadings. In the PCA, verbal abilities were positively loaded on PC2, whereas in the factor analysis, some verbal abilities have negative loadings on ML2.

Overall, both PCA and factor analysis results suggest the presence of a general cognitive factor. However, the second component (PC2) and factor highlight different aspects of the underlying structure in the data. PCA provides a purely statistical representation of variance without assuming an underlying model, while factor analysis attempts to uncover latent variables that might explain the observed correlations among the subtests.

Also, it is worth mentioning that the results of the factor analysis (in terms of how many factors are needed to explain the intercorrelations) match the conclusion in the PCA analysis. As shown by [Figure 4](#fig-scree-plot-02)

### d)

These are the communalities for the Information and Vocabulary subtests given by the model

round(fa\_02$communality[c(1,6)], 4)

Information Vocabulary   
 0.5261 0.6113

We can get the same result using the factor loading as follows:

round(c(sum(fa\_02$loadings[1,]\*\*2), sum(fa\_02$loadings[6,]\*\*2)), 4)

[1] 0.5261 0.6113

### e)

Based on the two factor model I computed, the Information and Comprehension subtests (let’s rename them and , respectively) can be express as follows:

Then, the correlation of this two variables can be compute as:

Using the assumptions of the two-factor model, we can reduce the latter expression into:

The estimations of these quantities can be found in the factor loadings:

fa\_02$loadings[c(1,2),c(1,2)]

ML1 ML2  
Information 0.6376020 -0.3457973  
Comprehension 0.5450952 -0.1819704

Hence, the fitted covariance (well, actually correlation) will be:

The observed covariance (correlation) between and is . This means that the residual for this correlation approx. will be:

Which can be obtain with the residual matrix:

fa\_02$residual[1,2]

[1] -0.04047867

This means that the model overestimated the correlation between the Information and Comprehension subtests (it was off by 0.04 units). However, the estimation yielded a plausible result.

# Question 3

### a)

In the one-factor model with three observed variables, the (symmetric) covariance matrix is the following:

Which means that we can get estimates for the parameters of the model by solving the following system of equations:

From the forth and fifth equation we have that:

We can plug-in these results into the sixth equation to get the following:

For the sake of this question, we would take the positive solution. This implies that:

Based on these estimations, we can also get the estimations for the :

### b)

The problem we would encounter is the overparameterization of the model. Since we have 3 observed variables (), the unique pieces of information (or degrees of freedom) available for the estimation of the model parameters is . However, a two-factor model requires at least degrees of freedom. This means that the two-factor model cannot be uniquely fitted to the covariance matrix because it does not have enough data points to estimate all the required parameters, aka the model is not identifiable.