Introduction

In this lab, we will perform various analysis under Factor Analysis. The dataset we will use is related to Salespeople data.

It has index of Sales Growth(**abbreviated with SG**), Sale Profitability (SP), New Account Sales(NS), Creativity Test (CT), Mechanical Reasoning(MR), Abstract Reasoning(AR) and Mathematics Test(MT) of 50 salesperson. We aim to find out relations between various variables in play among the 50 individuals, and might unfold hidden information about salesperson.

**Q1. Find the correlation matrix for the data, study it, and comment.**

We loaded the data of 50 individuals in R.

We can calculate the correlation between the variables using commands in R.

Refer Appendix 1 for Correlation Data generated.

From the same table, we can see that the highest correlation is seen between mathematics test and sales profitability, which shows that the sales individual if is good in Math tends to have higher sales profitability. We must note that this might not be a causation relationship, but just correlation.

Also, Sales growth and sales profitability are highly correlated implying that individuals that have sold more have higher profitability as well. It is obvious as more good sold will rise up sales profit naturally.

Least correlation is seen between abstract reasoning and creativity test, which is quite counter-intuitive. But, it also depends upon the type of creativity and abstract reasoning taken for the test.

Similarly, we can study correlation between various variables in play in this system.

**Q2. Check data normality (univariate and multivariate)**

We can check for univariate normality using QQ Plots. QQ plots plot the quantiles of a variable compared with theoretical quantiles and comes up with a plot that can reasonably conclude if the distribution of the variable is same as a theoretical normal distribution.

If the variable align with the central red line (mean line), then it means they are normal, else they are not normal.

We did the same for the seven variables under discussion.

We obtained the following results for the variables’ normality (**univariate Normality)**.

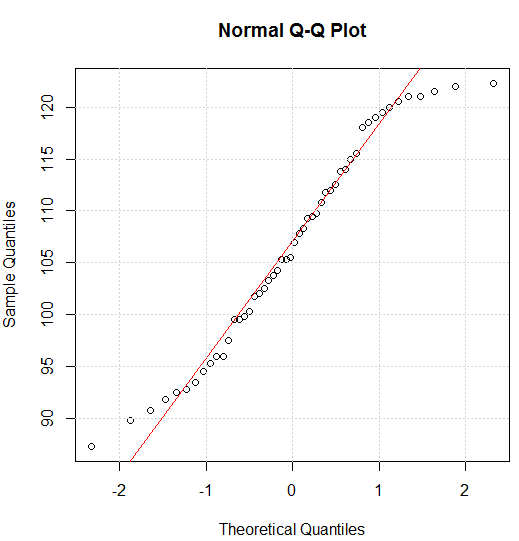


Fig2.2 Normal QQ Plot for Sales Profitability

We can see that the normal QQ plot for SP is almost normal. There are some individuals in top right and bottom left that look like outliers.

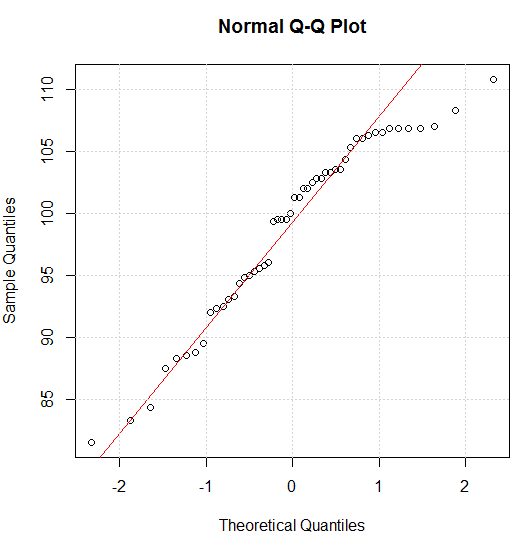


Fig 2.1 Normal QQ Plot for Sales Growth

We can see that the normal QQ plot for Sales Growth is almost normal. There are some individuals that are quite far from the mean line (seen in the top right corner)

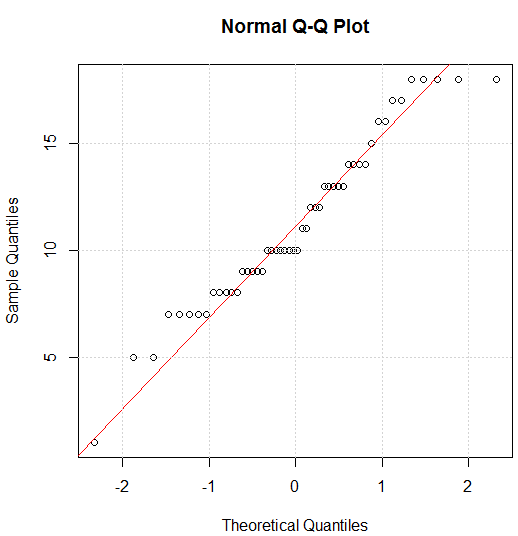


Fig 2.4 Normal QQ Plot for Creativity Test

These are somewhat normal.

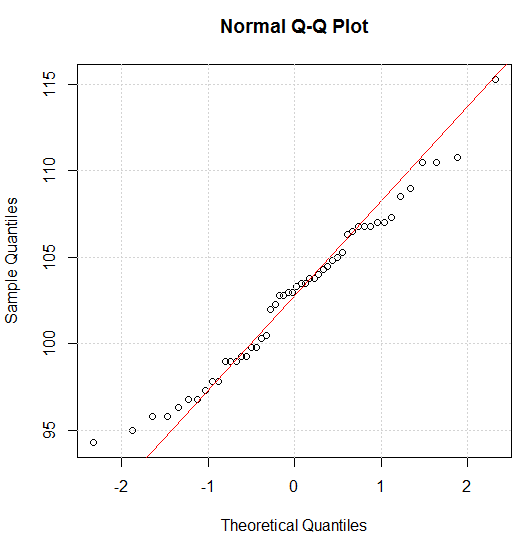


Fig 2.3 Normal QQ Plot for New Account Sales

We can see that the normal QQ plot for NS is almost normal.

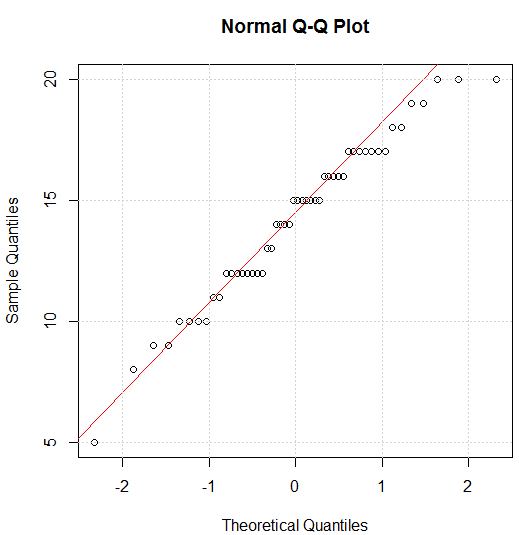


Fig 2.5 Normal QQ Plot for Mechanical Reasoning

These are somewhat normal.

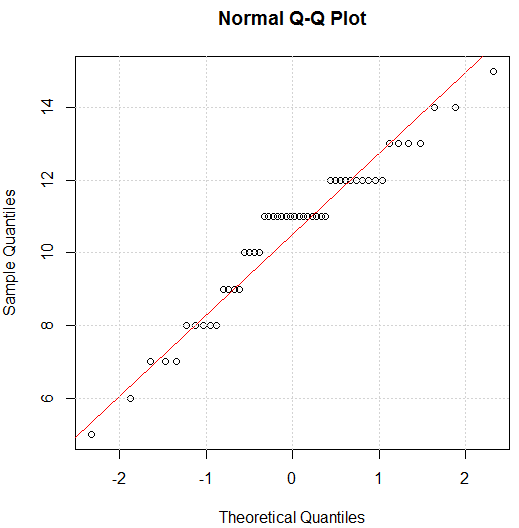


Fig 2.6 Normal QQ Plot for Abstract Reasoning

These are reasonably normal.

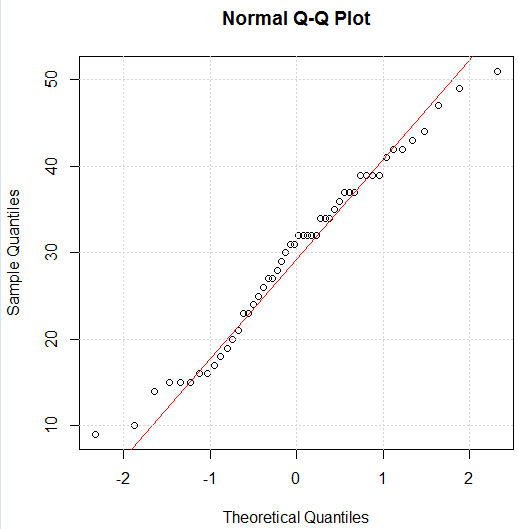


Fig 2.7 Normal QQ Plot for Mathematics

These are somewhat normal.

Thus, by using QQ plot we were able to find out that all the variables look almost normal: they were not randomly distributed and had reasonable degree of normality in them.

Now, let us test **multivariate normality** for our dataset.

We can perform multivariate normality analysis using Chi-square test.

We find out the inverse of Variance-Covariance matrix of the 7 variables, then calculate the distance of the 50 datasets using formula

D= XT ∑-1 X

And if this statistical distance follows chi-square distribution, then the dataset is multivariate-normal i.e. geometrically if the statistical distance from the mean is normally distributed, then the plot generated will coincide with the mean line, and hence the dataset is multi-normal.

In the adjacent figure, we can see that the QQ plot obtained for our data is not normal.

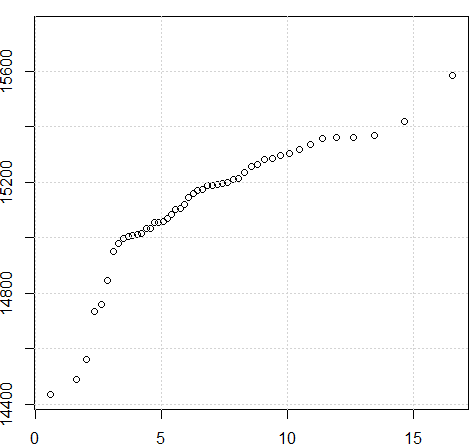


Fig 2.8 Chi-square test of multi-normality

**X-axis:** quantile function for chi-squared distribution with 7 degrees of freedom

**Y-axis:** statistical distance (d)

Hence, it does not satisfy multi-normality.

Also from Komogorov-Smirnov (KS) test, we have,

Distance (d)=1

P=.88e-16

Since, p is less than0 .05, this means that our null hypothesis (that the two samples were drawn from the same distribution) is invalid.

Also, distance which is the absolute max distance between CDFs of two samples has to be closer to 0, which is not the case.

Hence, our dataset does not satisfy multi-normality i.e. it is not normal.

**3) Perform FA with different number of common factors (try 1 to 5) and different rotations, analyze loadings and scores, find main groups of variables, find single-factor variables (if any), check scores’ normality. Find the optimal number of factors (consider the easiness of factor interpretation, number of factors, the proportion of the total variance explained, the residual matrix, etc.) Discuss and justify your choice. [This assignment will be evaluated on the basis of your discussion and justification. Support each of your statements by figures or numerical summaries of analysis.]**

**Factor Analysis:**

It is a statistical method used to describe the variability among observed variables in terms of potentially lower number of unobserved variables, called factors. The observed variables are modelled as linear combinations of the potential factors, plus “error” terms. By using factor analysis, we can reduce a dataset with large number of variables to one with relatively small number of variables.

If there are p random variable **x1,…. xp** with mean **μ1…. μp**. Then suppose for some unknown constants lij and k unobserved random variables **Fj** called common factors (because they influence all the observed random variables), where i=1,2,3,,,p and j= 1,2,……. k, such that k<p, we have

xi- μi = li1F1+ …+likFk + εi

In matrix terms, we can represent them as,

**x- μ = LF + ε**

where, L is matrix of factor loadings, F is factors, **ε** is the vector oferrors.

We also assume that F and ε are independent, expected value of F is zero and that all the factors are uncorrelated.

If cov(**x- μ)= Σ,** then

**Σ= LLT +Ψ**

and, Var(xi) = li12 + …. lim2 + **Ψi**

Communality Specific Variance

Finally, with this relation we will be able to explain the variability of our data in terms of factor loadings. **Ψ** is the covariance matrix of errors, which is a diagonal matrix.

**Factor Loadings**

The relationship of each variable to the underlying factor is expressed by factor loadings, represented by the loading matrix **L.**

One of the hardest things to do during Factor Analysis is how many factors to settle on. One of the way is to consider all those factors with **eigen value >=1**. This is because a factor with an eigenvalue of 1 accounts for as much variance as a single variable, and the logic is that only factors that explain at least the same amount of variance as a single variable is worth keeping.

But this should be taken only as a tool, not a hard and fast rule to select number of factors. Another option is using **Scree plot.** A scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. It always should be a downward curve. The point where the slope of the curve is leveling of indicated the number of factor that should be generated by the analysis. But it can also be just used as a tool rather than a hard and fast rule.

Another important metric is to keep in mind is **the total amount of variability** of the original variables explained by each variables explained by each factor solution.

We cannot be using more than three factors in this Factor Analysis, because from the formula:

M <= (p-1)/2,

Where M=no of factors

P=no of variables

For our case, p=7, therefore, m =1,2,3 .

Hence, we cannot work with greater than 3 number of factors.

Here we will now observe the value of loading, communalities and uniqueness of the different variable of our dataset, for different kinds of rotation and for different number of factors. By doing this brute force approach (and not using approaches of eigen values, scree plot) we are trying to select which model of FA would be best for understanding our data while looking at all of our possibilities.

Changing Rotation and number of factors for Bartlett Scoring.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Factor Loadings ( 1 Factor) | | | | |
| Variables | Unrotated | Varimax | Promax | Communalities | Uniqueness |
| Sales Growth | .975 | .975 | .975 | .95 | .05 |
| Sales Profitability | .959 | .959 | .959 | .92 | .08 |
| New Account Sales | .902 | .902 | .902 | .81 | .19 |
| Creativity Test | .567 | .567 | .567 | .32 | .68 |
| Mechanical Reasoning | .712 | .712 | .712 | .51 | .49 |
| Abstract Reasoning | .615 | .615 | .615 | .38 | .62 |
| Mathematic Tests | .953 | .953 | .953 | .91 | .09 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Factor Loadings (2 Factor)** | | | | | | | |
| **Variables** | Unrotated | | Varimax | | Promax | | Communalities | Uniqueness |
| **F1** | **F2** | **F1** | **F2** | **F1** | **F2** |
| Sales Growth | .695 | .669 | .852 | .452 | .896 | .109 | .93 | .07 |
| Sales Profitability | .669 | .695 | .868 | .419 | .926 |  | .93 | .07 |
| New Account Sales | .795 | .494 | .717 | .602 | .688 | .346 | .88 | .12 |
| Creativity Test | .983 | -.167 | .148 | .987 | -.116 | 1.063 | 1 | 0 |
| Mechanical Reasoning | .655 | .312 | .501 | .525 | .449 | .362 | .53 | .47 |
| Abstract Reasoning | .250 | .569 | .619 |  | .731 | -.230 | .39 | .61 |
| Mathematic Tests | .558 | .812 | .946 | .277 | 1.062 | -.138 | .98 | .02 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Factor Loadings (3 Factor)** | | | | | | | | | | |
| **Variables** | Unrotated | | | Varimax | | | Promax | | | Communalities | Uniqueness |
| **F1** | **F2** | **F3** | **F1** | **F2** | **F3** | **F1** | **F2** | **F3** |
| Sales Growth | 0.901 | 0.381 |  | 0.793 | 0.374 | 0.438 | 0.771 |  | 0.202 | 0.96 | 0.04 |
| Sales Profitability | .775 | 0.600 |  | 0.911 | 0.317 | 0.185 | 1.117 |  | -0.173 | 0.97 | 0.03 |
| New Account Sales | .931 | 0.202 |  | 0.651 | 0.544 | 0.438 | 0.457 | 0.380 | 0.261 | 0.91 | 0.09 |
| Creativity Test | .733 | -.118 | 0.666 | 0.255 | 0.964 |  | -0.168 | 1.143 | -0.109 | 1 | 0 |
| Mechanical Reasoning | .689 | .225 | 0.169 | 0.542 | 0.465 | 0.207 | 0.456 | .0339 |  | 0.55 | 0.45 |
| Abstract Reasoning | .757 | -.132 | -0.636 | 0.299 |  | 0.950 |  | -0.109 | 1.085 | 1 | 0 |
| Mathematic Tests | .762 | .608 | -0.110 | 0.917 | 0.18 | 0.298 | 1.145 | -0.225 |  | 0.96 | 0.04 |

In all these settings, we can see that the p-value from the chi square hypothesis test shows that none of the models fits the data perfectly since the p-value is very small.

It is worth mentioning that the best model if needed to be chosen in this setting would be Promax with 3 factors, as it explains for the 93.8% of the variance of all variables.

For this setting,

Factor 1 is positively related with Sales Growth, Sales Profitability, new account sales, mechanical reasoning and mathematics test, while it is almost similar to sales profitability and mathematics test, which might indicate that the hidden factor is Ability to sell Items.

Corelation Data for Qs 1.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Sales Growth | Sales Profiltability | New Account sales | Creativity test | Mechanical Reasoning | Abstract Reasoning | Mathematics Test |
| Sales Growth | 1 | 0.926076 | 0.884002 | 0.572036 | 0.708074 | 0.674407 | 0.927312 |
| Sales Profitability | 0.926076 | 1 | 0.842523 | 0.541508 | 0.74591 | 0.465388 | 0.944296 |
| New Account Sales | 0.884002 | 0.842523 | 1 | 0.700363 | 0.637471 | 0.641089 | 0.852568 |
| Creativity Test | 0.572036 | 0.541508 | 0.700363 | 1 | 0.590736 | 0.146907 | 0.41264 |
| Mechanical reasoning | 0.708074 | 0.74591 | 0.637471 | 0.590736 | 1 | 0.38595 | 0.574553 |
| Abstract Reasoning | 0.674407 | 0.465388 | 0.641089 | 0.146907 | 0.38595 | 1 | 0.566372 |
| Mathematics Test | 0.927312 | 0.944296 | 0.852568 | 0.41264 | 0.574553 | 0.566372 | 1 |

Table 1. Correlation Data for 7 variables of Dataset