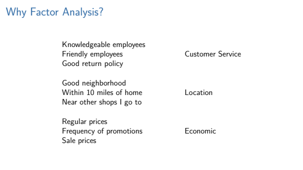
Factor Analysis

Why Factor Analysis?

Knowledgeable employees Friendly employees Good return policy	Customer Service
Good neighborhood Within 10 miles of home Near other shops I go to	Location
Regular prices Frequency of promotions Sale prices	Economic



Why Factor Analysis?

Knowledgeable employees Friendly employees Good return policy	Customer Service
Good neighborhood Within 10 miles of home Near other shops I go to	Location
Regular prices Frequency of promotions Sale prices	Economic

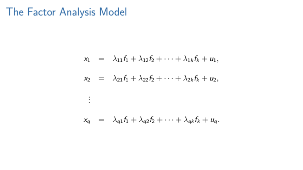
Exploratory factor analysis is a multivariate statistical procedure used to determine the underlying structure, in the form of latent - or unobserved - variables - among a larger set of observed variables. This can be thought of as a data reduction technique.

In this lesson, we will not be covering confirmatory factor analysis, which is similar, but would different in that the researcher would have some preconceived idea of what the underlying structure is and then try to confirm it.

In an exploratory factor analysis, we may have observable, measurable variables from customer surveys, like the things in the left-hand column here - Knowledgeable employees, Friendly employees, Good return policy, etc. These observed variables may

The Factor Analysis Model

$$
\begin{aligned}
x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + u_1, \\
x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_k + u_2, \\
&\vdots \\
x_q &= \lambda_{q1} f_1 + \lambda_{q2} f_2 + \cdots + \lambda_{qk} f_k + u_q.
\end{aligned}
$$



The Factor Analysis Model

$$\begin{aligned} x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + u_1, \\ x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_k + u_2, \\ &\vdots \\ x_q &= \lambda_{q1} f_1 + \lambda_{q2} f_2 + \cdots + \lambda_{qk} f_k + u_q. \end{aligned}$$

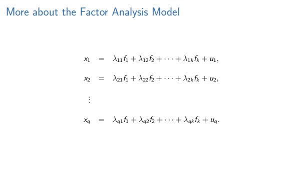
Meet the Factor Analysis model. This model is used when we believe that the observed variables, the x's, can be accurately represented by a smaller number of unobserved variables - represented here by the f's.

The f's are also called latent variables or "common factors" and there must be fewer f's than x's. The point is to reduce the number of variables and measure things that are thought to be not directly observable. Common examples of these factors are broad concepts like intelligence, social class, or burnout.

These factors are assumed to be independent and identically distributed and standardized with mean 0 and variance 1. Note that there is no requirement to be normally distributed, only identically

More about the Factor Analysis Model

$$\begin{aligned}
x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + u_1, \\
x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_k + u_2, \\
&\vdots \\
x_q &= \lambda_{q1} f_1 + \lambda_{q2} f_2 + \cdots + \lambda_{qk} f_k + u_q.
\end{aligned}$$



More about the Factor Analysis Model

$$\begin{aligned} x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + u_1, \\ x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_k + u_2, \\ &\vdots \\ x_q &= \lambda_{q1} f_1 + \lambda_{q2} f_2 + \cdots + \lambda_{qk} f_k + u_q. \end{aligned}$$

Let's continue talking about the specifics of this model. The lambda's are constants that are weights relating each observed variable x to the unobserved factors. The lambdas are called "factor loadings." The factor loadings represent the correlation between each variable and the factor.

The u's in the model are like residual terms, and are uncorrelated with each other and also uncorrelated with the factors f. Notice that there is a unique u corresponding to each observed variable x. Technically, they are called "specific variates".

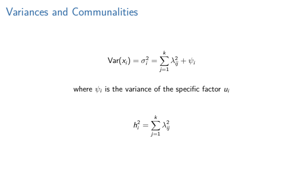
Since the u's are variables, they each have their own variance - which is called the "specific" or "unique" variance and represents the variability in each x that is NOT shared with other variables.

Variances and Communalities

$$\mathrm{Var}(x_i) = \sigma_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i$$

where ψ_i is the variance of the specific factor u_i

$$h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2$$



Variances and Communalities

$$\mathrm{Var}(x_i) = \sigma_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i$$

where ψ_i is the variance of the specific factor u_i

$$h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2$$

Since the factors are assumed to be standardized, each having a variance of 1, and they're independent, and since the lambda's are constants and the error terms are independent of the factors, the rules for variances would tell us that the variance for a particular x-sub-i is defined as shown here.

The variance of the specific factor u-sub-i, goes by the Greek letter psi and is called the "specific variance" or sometimes "unique variance." This is the variance that cannot be explained by the correlations to the other variables, but is still uniquely associated with a particular observed variable.

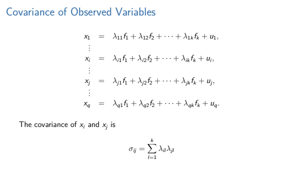
The sum of the squared factor loadings for a given variable x-sub-i represent the part of the overall variance known as the

Covariance of Observed Variables

$$\begin{aligned}
x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + u_1, \\
&\vdots \\
x_i &= \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{ik} f_k + u_i, \\
&\vdots \\
x_j &= \lambda_{j1} f_1 + \lambda_{j2} f_2 + \cdots + \lambda_{jk} f_k + u_j, \\
&\vdots \\
x_q &= \lambda_{q1} f_1 + \lambda_{q2} f_2 + \cdots + \lambda_{qk} f_k + u_q.
\end{aligned}$$

The covariance of x_i and x_j is

$$\sigma_{ij} = \sum_{l=1}^{k} \lambda_{il} \lambda_{jl}$$



Covariance of Observed Variables

$$\begin{aligned} x_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1k}f_k + u_1, \\ &\vdots \\ x_i &= \lambda_{i1}f_1 + \lambda_{i2}f_2 + \cdots + \lambda_{ik}f_k + u_i, \\ &\vdots \\ x_j &= \lambda_{j1}f_1 + \lambda_{j2}f_2 + \cdots + \lambda_{jk}f_k + u_j, \\ &\vdots \\ x_q &= \lambda_{q1}f_1 + \lambda_{q2}f_2 + \cdots + \lambda_{qk}f_k + u_q. \end{aligned}$$

The covariance of x_i and x_j is

$$\sigma_{ij} = \sum_{l=1}^{k} \lambda_{il}\lambda_{jl}$$

If all of the observed variables x-sub-i were independent, there would be no need to do a factor analysis. Its when there are groups of correlated x variables that we might start thinking that each group could be represented by an underlying, unobserved common factor.

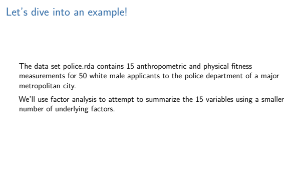
A correlation matrix can show us which x variables are correlated. Correlation is a scaled version of covariance, which is a measure of how two variables change together. The covariances for the x-sub-i's are given by the factor analysis model as the sum of the product of the factor loadings for each variable as shown here.

That is, the covariance of x-sub-i and x-sub-j is given by lambda-sub-i1 times lambda-sub-j1 plus lambda-sub-i 2 times lambda-sub-j 2, and so on through lambda-sub-i k times

Let's dive into an example!

The data set police.rda contains 15 anthropometric and physical fitness measurements for 50 white male applicants to the police department of a major metropolitan city.

We'll use factor analysis to attempt to summarize the 15 variables using a smaller number of underlying factors.



Let's dive into an example!

The data set police.rda contains 15 anthropometric and physical fitness measurements for 50 white male applicants to the police department of a major metropolitan city.

We'll use factor analysis to attempt to summarize the 15 variables using a smaller number of underlying factors.

bottom panel note: This data set is from Regression Analysis and its Application: A Data-Oriented Approach by Gunst and Mason (1980).

The Observed Variables

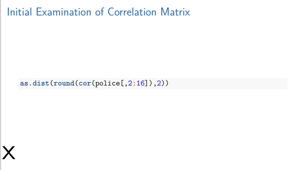
- REACT = Reaction time in seconds to a visual stimulus
- HEIGHT = Height in centimeters
- WEIGHT = Weight in kilograms
- SHLDR = Shoulder width in centimeters
- PELVIC = Pelvic width in centimeters
- CHEST = Minimum chest circumference in centimeters
- THIGH = Thigh skinfold thickness in millimeters
- PULSE = Resting pulse rate

The Observed Variables (cont'd)

- DIAST = Diastolic blood pressure
- CHNUP = Number of chin-ups the applicant was able to complete
- BREATH = Maximum breathing capacity in liters
- RECVR = Pulse rate after 5 minutes of recovery from treadmill running
- ENDUR = Treadmill endurance time in minutes
- SPEED = Maximum treadmill speed
- FAT = Total body fat measurement

Initial Examination of Correlation Matrix

```r
as.dist(round(cor(police[,2:16]),2))
```



Initial Examination of Correlation Matrix

```
as.dist(round(cor(police[,2:16]),2))
```

Our goal here is to determine if there is some underlying structure that can summarize the information found in these 15 variables using a smaller number of factors. Initially, we would like to see some degree of collinearity among these variables, unlike in multiple regression - where we don't want our predictor variables to be highly correlated with each other.

If there are clusters of variables that are correlated, then they have a shared variation that might be represented by a single factor.

An R code for getting this is shown here where some functions are layered to produce output that is a little easier to navigate, since 225 correlations will be produced in a 15 by 15 matrix.

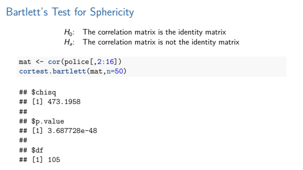
The function “as.dist” directs R to print only the lower left triangle

Bartlett's Test for Sphericity

H_0: The correlation matrix is the identity matrix
H_a: The correlation matrix is not the identity matrix

```
mat <- cor(police[,2:16])
cortest.bartlett(mat,n=50)
```

```
## $chisq
## [1] 473.1958
##
## $p.value
## [1] 3.687728e-48
##
## $df
## [1] 105
```



Bartlett's Test for Sphericity

H_0: The correlation matrix is the identity matrix
H_a: The correlation matrix is not the identity matrix

```
mat <- cor(police[,2:16])
cortest.bartlett(mat,n=50)
```

```
## $chisq
## [1] 473.1958
##
## $p.value
## [1] 3.687728e-48
##
## $df
## [1] 105
```

Bartlett's test for sphericity is one formal way to determine if factor analysis may be appropriate for our data set by testing that the observed correlation matrix is not an identity matrix. An identity matrix is a matrix with 1's on the diagonal and zeros everywhere else.

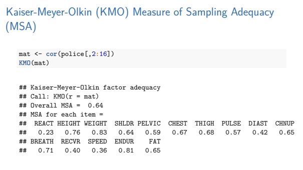
In other words, it tests that the population correlations among pairs of variables are not all zero. In this case, we see that the very small P-value gives ample evidence to reject the null hypothesis and conclude that our correlation matrix is not an identity matrix and that factor analysis may be useful here.

bottom panel note: The package "psych" must be installed and loaded to use the cortest.bartlett function.

Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA)

```r
mat <- cor(police[,2:16])
KMO(mat)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat)
## Overall MSA =  0.64
## MSA for each item =
##  REACT HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  DIAST  CHNUP
##   0.23   0.76   0.83   0.64   0.59   0.67   0.68   0.57   0.42   0.65
## BREATH  RECVR  SPEED  ENDUR    FAT
##   0.71   0.40   0.36   0.81   0.65
```



Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA)

```
mat <- cor(police[,2:16])
KMO(mat)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat)
## Overall MSA =  0.64
## MSA for each item =
##  REACT HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  DIAST  CHNUP
##   0.23   0.76   0.83   0.64   0.59   0.67   0.68   0.57   0.42   0.65
## BREATH  RECVR  SPEED  ENDUR    FAT
##   0.71   0.40   0.36   0.81   0.65
```

While Bartlett's test indicates only the presence of nonzero correlations, the KMO measure of sampling adequacy accounts for the patterns between variables in addition to the correlations.

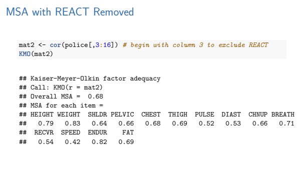
The KMO measure of sampling adequacy takes on values between 0 and 1. Values closer to 1 indicate that factor analysis should yield distinct and reliable factors. Kaiser (the K in KMO) in 1979 recommends values above .5 as acceptable to proceed with factor analysis.

Hutcheson and Sofroniou, in a 1999 publication conclude that MSA values between 0.5 and 0.7 are considered mediocre, values between 0.7 and 0.8 are considered good, values between 0.8 and 0.9 are deemed great and values above 0.9 are superb.

MSA with REACT Removed

```r
mat2 <- cor(police[,3:16]) # begin with column 3 to exclude REACT
KMO(mat2)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat2)
## Overall MSA =  0.68
## MSA for each item =
## HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  DIAST  CHNUP BREATH
##   0.79   0.83   0.64   0.66   0.68   0.69   0.52   0.53   0.66   0.71
##  RECVR  SPEED  ENDUR    FAT
##   0.54   0.42   0.82   0.69
```



MSA with REACT Removed

```
mat2 <- cor(police[,3:16]) # begin with column 3 to exclude REACT
KMO(mat2)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat2)
## Overall MSA =  0.68
## MSA for each item =
## HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  DIAST  CHNUP BREATH
##   0.79   0.83   0.64   0.66   0.68   0.69   0.52   0.53   0.66   0.71
##  RECVR  SPEED  ENDUR    FAT
##   0.54   0.42   0.82   0.69
```

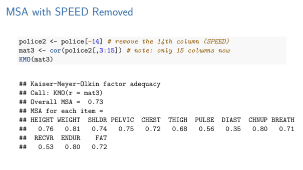
Notice now that the overall MSA is up to .68, still in the mediocre range but higher now and also only SPEED is below the .5 threshold. Therefore, SPEED will be removed and a new MSA will be computed.

bottom panel note: The package “psych” must be installed and loaded to use the KMO function.

MSA with SPEED Removed

```
police2 <- police[-14] # remove the 14th column (SPEED)
mat3 <- cor(police2[,3:15]) # note: only 15 columns now
KMO(mat3)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat3)
## Overall MSA =  0.73
## MSA for each item =
## HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  DIAST  CHNUP BREATH
##   0.76   0.81   0.74   0.75   0.72   0.68   0.56   0.35   0.80   0.71
##  RECVR  ENDUR    FAT
##   0.53   0.80   0.72
```



MSA with SPEED Removed

```
police2 <- police[-14] # remove the 14th column (SPEED)
mat3 <- cor(police2[,3:15]) # note: only 15 columns now
KMO(mat3)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat3)
## Overall MSA =  0.73
## MSA for each item = 
## HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  DIAST  CHNUP BREATH 
##   0.76   0.81   0.74   0.75   0.72   0.68   0.56   0.35   0.80   0.71 
##  RECVR  ENDUR    FAT 
##   0.53   0.80   0.72 
```

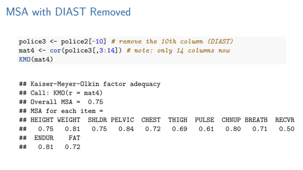
We see now an overall MSA of .73, which is deemed as a good measure, and now only diastolic blood pressure has individual MSA value under .5, so we'll remove that one and see what we get.

bottom panel note: The package "psych" must be installed and loaded to use the KMO function.

MSA with DIAST Removed

```
police3 <- police2[-10] # remove the 10th column (DIAST)
mat4 <- cor(police3[,3:14]) # note: only 14 columns now
KMO(mat4)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat4)
## Overall MSA =  0.75
## MSA for each item = 
## HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  CHNUP BREATH  RECVR 
##   0.75   0.81   0.75   0.84   0.72   0.69   0.61   0.80   0.71   0.50 
##  ENDUR    FAT 
##   0.81   0.72
```



MSA with DIAST Removed

```
police3 <- police2[-10] # remove the 10th column (DIAST)
mat4 <- cor(police3[,3:14]) # note: only 14 columns now
KMO(mat4)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat4)
## Overall MSA =  0.75
## MSA for each item =
## HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  CHNUP BREATH  RECVR
##   0.75   0.81   0.75   0.84   0.72   0.69   0.61   0.80   0.71   0.50
##  ENDUR    FAT
##   0.81   0.72
```

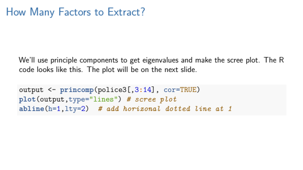
We see now an overall MSA of .75, a good measure, and now no individual MSA values are under .5, though pulse after 5 minutes recovery from treadmill run is right on it. For completeness, we could conduct Bartlett's test again on the reduced data set. I'll skip displaying it for you here, but rest assured that the p-value was exceedingly small and we will proceed with the factor analysis.

bottom panel note: The package "psych" must be installed and loaded to use the KMO function.

How Many Factors to Extract?

We'll use principle components to get eigenvalues and make the scree plot. The R code looks like this. The plot will be on the next slide.

```r
output <- princomp(police3[,3:14], cor=TRUE)
plot(output,type="lines") # scree plot
abline(h=1,lty=2)  # add horizonal dotted line at 1
```



How Many Factors to Extract?

We'll use principle components to get eigenvalues and make the scree plot. The R code looks like this. The plot will be on the next slide.

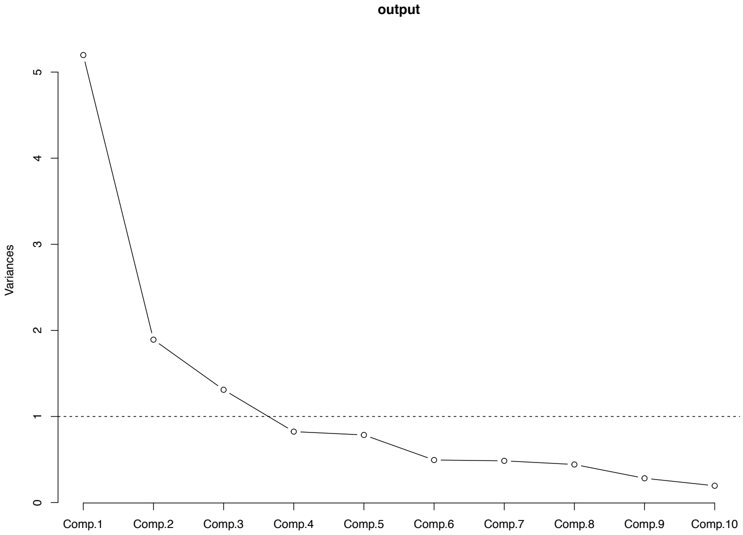
```
output <- princomp(police3[,3:14], cor=TRUE)
plot(output,type="lines") # scree plot
abline(h=1,lty=2)  # add horizonal dotted line at 1
```

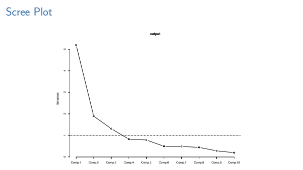
Sometimes it is known or presumed how many latent factors underly the structure of the data and so we can specify directly how many factors to extract. Other times, we must rely on various techniques for guiding us to the best number of factors to extract.

One common approach is to extract factors for which the latent root is greater than 1 - this is called Kaiser's Rule. The latent root is also called the eigenvalue, which is the sum of squared loadings on that factor.

Another common approach is the scree test criterion. The scree plot contains the latent roots in the vertical axis and the number of factors in ther order of extraction on the horizontal axis. These latent roots are also called eigenvalues and R labels them as SS

Scree Plot


output
Variances
0
1
2
3
4
5
Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
Comp.6
Comp.7
Comp.8
Comp.9
Comp.10



Scree Plot

The recommended number of factors to extract is determined by a visual examination of where the plot begins to straighten out. Right here in this plot. This is called the “knee” in the plot. According to the scree test criterion, we should extract 3 factors, since the plot somewhat straightens out at 3 factors. This is a subjective judgement.

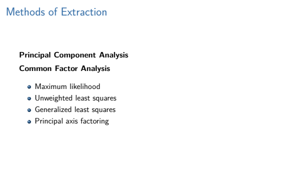
The scree plot is shown here with a dotted horizontal line at 1 and we can see that there are 4 factors with latent roots (eigenvalues) above 1. This criterion is thought to be useable whan there are fewer than 30 variables. The latent roots/eigenvalues here are labed automatically by the princomp function output as “variances.”

Methods of Extraction

Principal Component Analysis

Common Factor Analysis

- Maximum likelihood
- Unweighted least squares
- Generalized least squares
- Principal axis factoring



Methods of Extraction

Principal Component Analysis

Common Factor Analysis

- Maximum likelihood
- Unweighted least squares
- Generalized least squares
- Principal axis factoring

We've finally come to the point of extracting the factors, but there are more decisions to make. Which extraction method shall we use? Should we use a factor rotation or not? If so, which one?

Principal components, or PCA, extraction forms uncorrelated linear combinations of the observed variables. The first component accounts for the most variance and successive components explain progressively smaller portions of the variance.

With principal components extraction, the minimum number of factors is sought out to account for the maximum amount of the total variation in the observed variables. It is recommended when the goal is data reduction and is probably the way to go in data science applications.

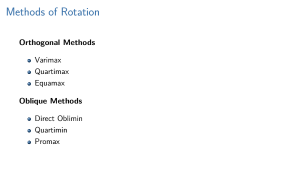
Methods of Rotation

Orthogonal Methods

- Varimax
- Quartimax
- Equamax

Oblique Methods

- Direct Oblimin
- Quartimin
- Promax



Methods of Rotation

Orthogonal Methods

- Varimax
- Quartimax
- Equamax

Oblique Methods

- Direct Oblimin
- Quartimin
- Promax

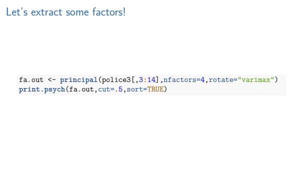
The term rotation is referring to turning the axes of the factors about the origin to some other position. Rotating factors is done primarily to attempt to simplfy the interpretation of the factors.

Orthogonal methods produce factors that are uncorrelated - that is , an angle of 90 degrees is maintained between the reference axes, while Oblique methods do not insist on this restriction.

The Varimax rotation tends to be commonly used.

Let's extract some factors!

```r
fa.out <- principal(police3[,3:14],nfactors=4,rotate="varimax")
print.psych(fa.out,cut=.5,sort=TRUE)
```



Let's extract some factors!

```
fa.out <- principal(police3[,3:14],nfactors=4,rotate="varimax")
print.psych(fa.out,cut=.5,sort=TRUE)
```

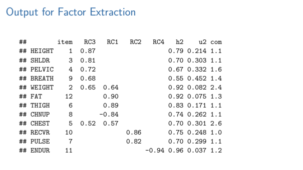
I've requested the Varimax rotation here, but even if no rotation is specified, Varimax is the dafault for the rotation in R, so if you don't want a rotation, you have to specify that by entering rotation="none".

Specifying cut at .5 suppresses the output for all of the factor loadings under .5 and sort=TRUE will sort the displayed loadings. This makes the output much easier to read.

The relevant output is on the next slide.

Output for Factor Extraction

```
##        item   RC3   RC1   RC2   RC4   h2    u2 com
## HEIGHT    1  0.87                   0.79 0.214 1.1
## SHLDR     3  0.81                   0.70 0.303 1.1
## PELVIC    4  0.72                   0.67 0.332 1.6
## BREATH    9  0.68                   0.55 0.452 1.4
## WEIGHT    2  0.65  0.64             0.92 0.082 2.4
## FAT      12        0.90             0.92 0.075 1.3
## THIGH     6        0.89             0.83 0.171 1.1
## CHNUP     8       -0.84             0.74 0.262 1.1
## CHEST     5  0.52  0.57             0.70 0.301 2.6
## RECVR    10              0.86       0.75 0.248 1.0
## PULSE     7              0.82       0.70 0.299 1.1
## ENDUR    11                   -0.94 0.96 0.037 1.2
```



Output for Factor Extraction

```
##        item   RC3   RC1   RC2   RC4   h2    u2 com
## HEIGHT    1  0.87                   0.79 0.214 1.1
## SHLDR     3  0.81                   0.70 0.303 1.1
## PELVIC    4  0.72                   0.67 0.332 1.6
## BREATH    9  0.68                   0.55 0.452 1.4
## WEIGHT    2  0.65  0.64             0.92 0.082 2.4
## FAT      12        0.90             0.92 0.075 1.3
## THIGH     6        0.89             0.83 0.171 1.1
## CHNUP     8       -0.84             0.74 0.262 1.1
## CHEST     5  0.52  0.57             0.70 0.301 2.6
## RECVR    10              0.86       0.75 0.248 1.0
## PULSE     7              0.82       0.70 0.299 1.1
## ENDUR    11                   -0.94 0.96 0.037 1.2
```

The communalities are in the h2 column and uniquenesses are in the the column labeled as u2. Remember that the communality is the proportion of common or shared variance within a variable. For a successful factor analysis, we would like to see a good number of higher values in the h2 column.

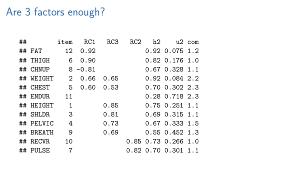
We see here that 11 of the 12 variables have communalities above 0.6 when 4 factors are extracted, which is very good.

Some suggest that variables with communalities less than .5 may be dropped from the factor analysis.

Note that the unique or specific variance, given in the u2 column, is simply 1 minus the communality for each variable, since the variables are all scaled to have a variance of 1 and total variance is

Are 3 factors enough?

```
##        item   RC1   RC3   RC2   h2    u2 com
## FAT      12  0.92             0.92 0.075 1.2
## THIGH     6  0.90             0.82 0.176 1.0
## CHNUP     8 -0.81             0.67 0.328 1.1
## WEIGHT    2  0.66  0.65       0.92 0.084 2.2
## CHEST     5  0.60  0.53       0.70 0.302 2.3
## ENDUR    11                   0.28 0.718 2.3
## HEIGHT    1        0.85       0.75 0.251 1.1
## SHLDR     3        0.81       0.69 0.315 1.1
## PELVIC    4        0.73       0.67 0.333 1.5
## BREATH    9        0.69       0.55 0.452 1.3
## RECVR    10              0.85 0.73 0.266 1.0
## PULSE     7              0.82 0.70 0.301 1.1
```



Are 3 factors enough?

```
##         item   RC1   RC3   RC2   h2    u2 com
## FAT      12  0.92              0.92 0.075 1.2
## THIGH     6  0.90              0.82 0.176 1.0
## CHNUP     8 -0.81              0.67 0.328 1.1
## WEIGHT    2  0.66  0.65        0.92 0.084 2.2
## CHEST     5  0.60  0.53        0.70 0.302 2.3
## ENDUR    11                    0.28 0.718 2.3
## HEIGHT    1        0.85        0.75 0.251 1.1
## SHLDR     3        0.81        0.69 0.315 1.1
## PELVIC    4        0.73        0.67 0.333 1.5
## BREATH    9        0.69        0.55 0.452 1.3
## RECVR    10              0.85 0.73 0.266 1.0
## PULSE     7              0.82 0.70 0.301 1.1
```

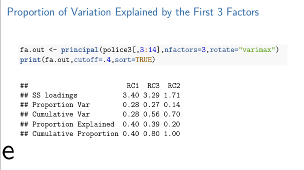
Again we're seeing 11 out of 12 variables with communalities aver .5, which is a good sign for extracting 3 factors and the one below .5 is the treadmill endurance, which we saw previously could stand on its own. We'll look at more of the output for this extraction on the next slide - examining the cumulative proportion of variance explained by the 3 factors.

Proportion of Variation Explained by the First 3 Factors

```
fa.out <- principal(police3[,3:14],nfactors=3,rotate="varimax")
print(fa.out,cutoff=.4,sort=TRUE)
```

```
##                         RC1  RC3  RC2
## SS loadings           3.40 3.29 1.71
## Proportion Var        0.28 0.27 0.14
## Cumulative Var        0.28 0.56 0.70
## Proportion Explained  0.40 0.39 0.20
## Cumulative Proportion 0.40 0.80 1.00
```

└Proportion of Variation Explained by the First 3 Factors



Proportion of Variation Explained by the First 3 Factors

```r
fa.out <- principal(police3[,3:14],nfactors=3,rotate="varimax")
print(fa.out,cutoff=.4,sort=TRUE)
```

```
##                        RC1  RC3  RC2
## SS loadings           3.40 3.29 1.71
## Proportion Var        0.28 0.27 0.14
## Cumulative Var        0.28 0.56 0.70
## Proportion Explained  0.40 0.39 0.20
## Cumulative Proportion 0.40 0.80 1.00
```

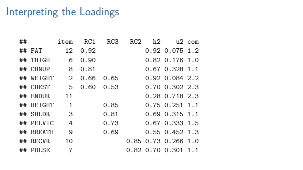
I'm displaying the portion of output that shows the sums of square for the loadings (also called the eigenvalues or the latent roots) as well as the proportion of total variance each factor explains and the cumulative proportion of variance the factors explain in the observed variables.

Notice that the first 3 factors explain a total of 70% of the total variation among the 12 remaining observed variables.

Recall that I said previously that one criterion for deciding how many factors to extract was the total proportion of variance explained by them and that often 60% was a minimum cutoff; here we are getting 70%.

Interpreting the Loadings

```
##         item   RC1   RC3   RC2   h2    u2 com
## FAT       12  0.92             0.92 0.075 1.2
## THIGH      6  0.90             0.82 0.176 1.0
## CHNUP      8 -0.81             0.67 0.328 1.1
## WEIGHT     2  0.66  0.65       0.92 0.084 2.2
## CHEST      5  0.60  0.53       0.70 0.302 2.3
## ENDUR     11                   0.28 0.718 2.3
## HEIGHT     1        0.85       0.75 0.251 1.1
## SHLDR      3        0.81       0.69 0.315 1.1
## PELVIC     4        0.73       0.67 0.333 1.5
## BREATH     9        0.69       0.55 0.452 1.3
## RECVR     10              0.85 0.73 0.266 1.0
## PULSE      7              0.82 0.70 0.301 1.1
```



Interpreting the Loadings

```
##          item   RC1   RC3   RC2   h2    u2 com
## FAT        12  0.92             0.92 0.075 1.2
## THIGH       6  0.90             0.82 0.176 1.0
## CHNUP       8 -0.81             0.67 0.328 1.1
## WEIGHT      2  0.66  0.65       0.92 0.084 2.2
## CHEST       5  0.60  0.53       0.70 0.302 2.3
## ENDUR      11                   0.28 0.718 2.3
## HEIGHT      1        0.85       0.75 0.251 1.1
## SHLDR       3        0.81       0.69 0.315 1.1
## PELVIC      4        0.73       0.67 0.333 1.5
## BREATH      9        0.69       0.55 0.452 1.3
## RECVR      10              0.85 0.73 0.266 1.0
## PULSE       7              0.82 0.70 0.301 1.1
```

Remember that factor loadings represent correlations between the original variables and the common factors.

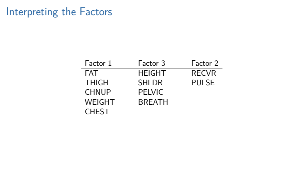
The thing to look for here is which variables load high, like more than .5 or .6 in absolute value on each factor. I have suppressed loadings under .5 in absolute value in the output and sorted the loadings to make it easier to see the higher loadings and which factors they load onto.

Starting with factor 1 (labeled here as PC1), we see that FAT, THIGH, CHNUP, WEIGHT, and CHEST load highly, and all of them load positively except for CHNUP, which is negatively correlated with factor 1.

For factor 3 (labeled here as PC3), we have HEIGHT, SHLDR,

Interpreting the Factors

Factor 1	Factor 3	Factor 2
FAT	HEIGHT	RECVR
THIGH	SHLDR	PULSE
CHNUP	PELVIC	
WEIGHT	BREATH	
CHEST		



Interpreting the Factors

Factor 1	Factor 3	Factor 2
FAT	HEIGHT	RECVR
THIGH	SHLDR	PULSE
CHNUP	PELVIC	
WEIGHT	BREATH	
CHEST		

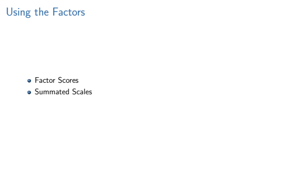
Another reason I wanted to go with 3 factors instead of 4 or 5 is that I was thinking ahead to this time, when I need to try to describe the underlying factors that the observed variables are representing. These physical characteristics are all very similar in many respects.

Interpreting factors is generally much better facilitated by a subject area expert, but I'll give it a shot. Factor 1 may be physical characteristics related to being overweight - particularly with very high positive loadings on total body fat, thigh skinfold, a high negative loading on the number of chin-ups the applicant completed, a pretty high positive loading on weight, and higher positive load on chest . Maybe we could use the label of "obesity" for factor 1.

Factor 3 may be some measure of skeletal structure with variables

Using the Factors

- Factor Scores
- Summated Scales



Using the Factors

- Factor Scores
- Summated Scales

Factor analysis is about more than just identifying the latent factors. We would like to make use of these factors, typically in some other statistical model or application. Maybe a multiple regression or ANOVA, for example.

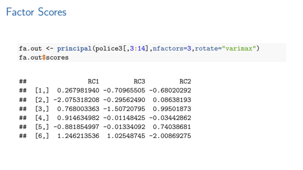
Since we haven't observed or measured these factors directly, we need some way to quantify what they represent. One option is through factor scores estimated from the factor anaolysis and the other is simply to combine the values of the variables that load highly onto each factor - typically by using the mean of them for each individual.

A factor score is a composite measure based on all of the factor loadings for each factor, of course giving more weight to variables

Factor Scores

```
fa.out <- principal(police3[,3:14],nfactors=3,rotate="varimax")
fa.out$scores
```

```
##                RC1         RC3         RC2
##  [1,]  0.267981940 -0.70965505 -0.68020292
##  [2,] -2.075318208 -0.29562490  0.08638193
##  [3,]  0.768003363 -1.50720795  0.99501873
##  [4,]  0.914634982 -0.01148425 -0.03442862
##  [5,] -0.881854997 -0.01334092  0.74038681
##  [6,]  1.246213536  1.02548745 -2.00869275
```



Factor Scores

```
fa.out <- principal(police3[,3:14],nfactors=3,rotate="varimax")
fa.out$scores
```

```
##                RC1         RC3         RC2
## [1,]  0.267981940 -0.70965505 -0.68020292
## [2,] -2.075318208 -0.29562490  0.08638193
## [3,]  0.768003363 -1.50720795  0.99501873
## [4,]  0.914634982 -0.01148425 -0.03442862
## [5,] -0.881854997 -0.01334092  0.74038681
## [6,]  1.246213536  1.02548745 -2.00869275
```

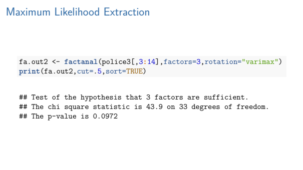
Factor scores are computed by the principal function and stored in an output object, if one is assigned. In the code shown here the output object is called fa.out. The summated scores are automatically assigned the name scores.

I'm only showing the first 6 scores for each factor, but there is a factor score for each of the 50 individuals in the sample corresponding to each of the 3 factors.

Maximum Likelihood Extraction

```
fa.out2 <- factanal(police3[,3:14],factors=3,rotation="varimax")
print(fa.out2,cut=.5,sort=TRUE)
```

```
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 43.9 on 33 degrees of freedom.
## The p-value is 0.0972
```



Maximum Likelihood Extraction

```
fa.out2 <- factanal(police3[,3:14],factors=3,rotation="varimax")
print(fa.out2,cut=.5,sort=TRUE)
```

```
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 43.9 on 33 degrees of freedom.
## The p-value is 0.0972
```

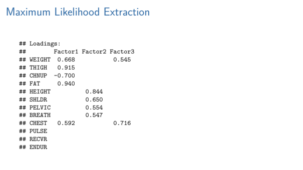
With maximum likelihood extraction, we're under the restriction that we must assume a multivariate normal structure to the variables, but one of the benefits it comes with is a hypothesis test for the number of factors to extract.

Notice the difference in the R code. The function factanal uses maximum likelihood as the extraction method, the data frame is given in the same way as with the principal function, the number of factors to extract goes by the word factors rather than nfactors, and the various rotation methods can be specified in the same way although with the principal function you can say either rotate or rotation, but factanal only accepts rotation.

I'm only displaying a portion of the output in this slide - just the

Maximum Likelihood Extraction

```
## Loadings:
##         Factor1 Factor2 Factor3
## WEIGHT  0.668           0.545
## THIGH   0.915
## CHNUP  -0.700
## FAT     0.940
## HEIGHT          0.844
## SHLDR           0.650
## PELVIC          0.554
## BREATH          0.547
## CHEST   0.592           0.716
## PULSE
## RECVR
## ENDUR
```



Maximum Likelihood Extraction

```
## Loadings:
##        Factor1 Factor2 Factor3
## WEIGHT  0.668           0.545
## THIGH   0.915
## CHNUP  -0.700
## FAT     0.940
## HEIGHT          0.844
## SHLDR           0.650
## PELVIC          0.554
## BREATH          0.547
## CHEST   0.592           0.716
## PULSE
## RECVR
## ENDUR
```

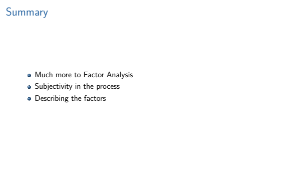
Notice a slightly different solution with maximum likelihood extraction, even using the Varimax rotation as before.

Factor 1 is the same as with principal components extraction with high positive loadings on total body fat, thigh skinfold, a high negative loading on the number of chin-ups the applicant completed, a pretty high positive loading on weight, and higher positive load on chest, though it is cross-loaded.

The same variablesthat loaded highly onto Factor 3 before is called Factor 2 here, but accounts for the second highest amount of variability. These are height, shoulder width, and pelvic width, and maximum breath capacity.
Weight is also cross-loaded and the cross-loaded chest loads higher

Summary

- Much more to Factor Analysis
- Subjectivity in the process
- Describing the factors



Summary

- Much more to Factor Analysis
- Subjectivity in the process
- Describing the factors

We're really just scratching the surface here in this presentation of Factor Analysis. There are many nuance issues that we haven't discussed and a large number of subjective decisions that need to be made when doing Factor Analysis.

Sometimes the latent factors will be more identifiable and other times it will be difficult to put a label on them. But is it kind of fun!