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

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

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

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

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

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

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

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.

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

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

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
```

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
```

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 SPEED Removed

```r
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 DIAST Removed

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

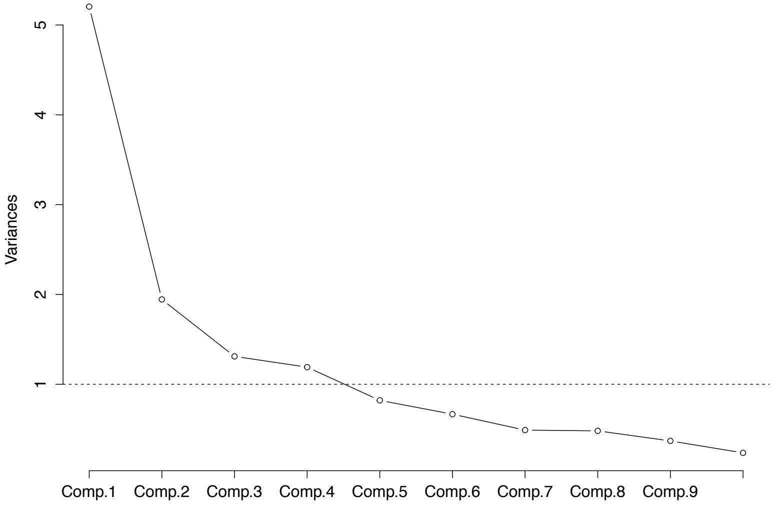
```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = mat4)
## Overall MSA =  0.75
## MSA for each item =
## HEIGHT WEIGHT  SHLDR PELVIC  CHEST  THIGH  PULSE  CHNUP BREATH  RECVR
##   0.75   0.81   0.75   0.84   0.72   0.69   0.61   0.80   0.71   0.50
##  ENDUR    FAT
##   0.81   0.72
```

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
```

Scree Plot



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

Methods of Extraction

Principal Component Analysis

Common Factor Analysis

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

Methods of Rotation

Orthogonal Methods

- Varimax
- Quartimax
- Equamax

Oblique Methods

- Direct Oblimin
- Quartimin
- Promax

Let's extract some factors!

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

Output for Factor Extraction

```
##          item  PC3   PC1   PC2   PC4   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
```

Are 3 factors enough?

```
##          item   PC1   PC3   PC2   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
```

Proportion of Variation Explained by the First 3 Factors

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

```
##                        PC1  PC3  PC2
## 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
```

Interpreting the Loadings

```
##           item   PC1   PC3   PC2   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 Factors

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

Using the Factors

- Factor Scores
- Summated Scales

Factor Scores

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

```
##              PC1         PC3         PC2
## [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
```

Maximum Likelihood Extraction

```r
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

```
## 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
```

Summary

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