STAT 8250 — Applied Multivariate Methods
Lab – Due Monday, Nov. 20

Example:

The following example is adapted from Johnson (*Applied Multivariate Methods for Data Analysts*, ch. 6). See this reference (on reserve) for more details.

The table below contains a partial listing of a data set consisting of anthropometric and physical fitness measurements that were taken on 50 white male applicants to the police department of a major metropolitan city. The variables include reaction time in seconds to a visual stimulus (REACT), the applicant's height in centimeters (HEIGHT), the applicant's weight in kilograms (WEIGHT), the applicant's shoulder width in centimeters (SHLDR), the applicant's pelvic width in centimeters (PELVIC), the applicant's minimum chest circumference in centimeters (CHEST), the applicant's thigh skinfold thickness in millimeters (THIGH), the applicant's resting pulse rate (PULSE), the applicant's diastolic blood pressure (DIAST), the number of chin-ups the applicant was able to complete (CHNUP), the applicant's maximum breathing capacity in liters (BREATH), the applicant's pulse rate after 5 minutes of recovery from treadmill running (RECVR), the applicant's maximum treadmill speed (SPEED), the applicant's treadmill endurance time in minutes (ENDUR), and the applicant's total body fat measurement (FAT).

Copy the file police1.sas from the data directory or from the web, run it, and examine the program and its associated output. Police1.sas uses SAS' PROC FACTOR to perform a factor analysis of the police applicant data.

The first step is to perform a PCA on the correlation matrix of the 15 variables in the data set. This could be done in PROC PRINCOMP, but PROC FACTOR will also perform a PCA, and it has an advantage in that it will automatically produce a scree plot when the SCREE option is specified on the PROC FACTOR statement. An examination of the scree plot on p.2 of police1.lst does not provide an indication of the dimensionality of the data that is as clear-cut as would be desirable. However, there are breaks after the second and fifth eigenvalues, and 5 PCs account for 75.8% of the total variability in the data. The proper number of factors in a FA does not necessarily equal the number of PCs that we would select in a PCA, but as first guess, $m = 5$ based on the PCA is a good starting point.

As the next step we fit the FA model based on $m = 5$ factors using the ML method. As an alternative to the ML method, or better yet, in addition to the ML method, we could try using the PC method to fit the $m = 5$-factor model and FA models at whatever other values of m that we decide to consider, but in this analysis we limit attention to result of the ML method. In addition to the METHOD=ML option, a varimax rotation was requested by ROTATE=VARIMAX, option S requests simple statistics (means, SDs), option C requests the correlation matrix, option EV requests that PROC FACTOR print the eigenvectors of the matrix being analyzed (the sample covariance matrix, in this case), the RES option requests the residual correlation and partial correlation matrices (discussed below), the REORDER option reorders the variables when printing out the estimated factor loading matrix and other matrices in the output so that the variables with the highest absolute loading on the first factor are displayed first, followed by variables with their highest absolute loading on the second factor, and so on (this just makes it a little easier to pick out which variables are most highly correlated with which factors, which is necessary at the factor interpretation stage). Finally, the option SCORE produces the matrix $\mathbf{S}^{-1}\hat{\mathbf{L}}$. This matrix's transpose can be used to compute factor scores via the regression method using formula (9-58), p.554 of our text. The factor scores themselves are computed and placed in data set SCORES specified in the OUT=SCORES option.

On pp.4–5 of police1.lst, PROC FACTOR produces the means and SDs of the 15 variables and their correlation matrix $\mathbf{R}$. The ML method is based on maximizing a loglikelihood function of the parameters of the FA model. This is done using an iterative numerical method that requires starting values for the parameters which are updated, and re-updated until convergence at the final estimates. On the top of p.6 are printed the starting values for the communalities based on the default method in which the communality for a variable is set equal to the squared multiple correlation (SMC) of that variable with all other variables. The iteration history is printed on the bottom of p.6. Notice that during the second iteration, the ML method ran into a problem in that a communality estimate was updated to a value greater than 1.0. Since communalities are by definition ≤ 1, the program produce an error message and stopped.

SAS has a HEYWOOD option that deals with this situation (known as a Heywood case) by taking any communality estimate greater than 1 encountered in the ML iterations and setting them to 1 before continuing the iterations. The third call to PROC FACTOR in police1.sas makes use of this option in refitting the $m = 5$-factor model with ML.

Notice that now on iteration 2 (see p.9 of police1.lst), the communality that had been > 1 (1.02633, for RECVR) is set to 1.0. The fitting routine converged after 10 iterations (p.10) to final ML estimates. The test on p.10 for H_0 : No common factors, is the test of independence that we discussed earlier in the course (see pp.122–124 of the class notes). The LRT of H_0 : 5 factors are sufficient, yields $p = .0714$, so we would not reject the adequacy of the 5-factor model at significance level .05. That is, it appears that increasing the number of factors is not necessary, although it may be possible to decrease m without significantly worsening the fit of the model. Also on p.10 the values for the two model selection criteria AIC and SBC are printed. These become meaningful when compared to the corresponding values produced for models with different values of m.

The matrix labelled FACTOR PATTERN on pp.11–12 is just $\hat{\mathbf{L}}$, the estimated factor loading matrix. Prior to rotation, it is not usually worthwhile to try to interpret

these loadings. The matrix on the bottom of p.12 and top of p.13 is $\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}'$. Its diagonal elements are the estimated specific variances and the off-diagonal elements are the differences between the true sample correlations and their estimates according to the 5−factor model. Obviously, if these values are close to zero, then we would take this as evidence that the model fits well. To summarize the size of all of these off-diagonal elements, PROC FACTOR computes the root mean square (a measure of size, irrespective of sign) overall and by variable. Again, we'd like these root mean square values to be small, and if any particular variable has a large root mean square residual value then that gives some indication of which variables aren't adequately accounted for in the current FA model. In this case, all of the root mean square values appear to be small enough to support the adequacy of the model. (How small these values should be is somewhat subjective and depends upon the strength of the correlations among the variables that we started with. However, typically we would not want to see any residual correlations $> .40$ and only one or two greater than .25.)

Another matrix that gives information on the quality of the fit of the FA solution is the partial correlation matrix appearing on p.14. This matrix gives the partial correlations among the variables, after controlling for the influence of the factors in the model. E.g., the partial correlation between REACT and HEIGHT is the correlation that you would obtain if you first regressed REACT on the five factors in the FA model and obtained the residuals, regressed HEIGHT on those five factors and obtained the residuals, and then computed the correlation between those two sets of residuals. It is the correlation between REACT and HEIGHT after removing the contribution from the factors to this correlation. Again, the root mean square values at the top of p.15 computed from this matrix should be small if the model fits well. In this case, the evidence seems again to be that the model fits adequately.

Near the middle of p.16 is the varimax-rotated matrix of factor loadings. (The orthogonal matrix used to obtain this rotation is at the top of p.16, but there is little to be gained in examining this transformation matrix.) The rotated matrix of loadings is the main result of fitting the FA model, and should be interpreted. The loadings themselves have interpretations as the estimated correlation between the original variables and the factors. Variables that are highly correlated (absolute correlations that are greater than .4) with the five rotated factor axes and their correlations are as follows:

Factor 1: THIGH (.96), FAT (.90), CHNUP (-.69), WEIGHT (.61), and CHEST (.49)

Factor 2: HEIGHT (.89), SHLDR (.75), WEIGHT (.62), BREATH (.60), PELVIC (.59) and CHEST (.46)

Factor 3: RECVR (.95), PULSE (.57), and SPEED (-.53)

Factor 4: WEIGHT (.42), and CHEST (.66)

Factor 5: REACT (.78)

Interpretation of the rotated factors requires a mix of knowledge, experience, discretion, wisdom and objectivity. It should be kept in mind that the underlying factors that one is trying to identify and describe must be unique and independent characteristics of the population represented in the data set being analyzed.

From the correlations listed above it is clear that factors 1,2, and 4 each have something to do with body size. But how do they differ? Keeping in mind that these three aspects of body size must be independent, it seems that the first is a measure of body mass or lack of it; that is, it is a measure of obesity. The second factor appears to be a

measure of frame size or skeletal structure. An interpretation for the fourth factor is more difficult, but perhaps this factor represents upper body strength, in particular it may be related to the extent to which the applicant lifts weights for strength conditioning. The third factor might be interpreted as a measure of aerobic or cardiovascular fitness. The fifth factor is what is known as a trivial factor, that is, a factor that is highly correlated with only one variable. In this case, the fifth factor seems to be a measurement of reaction time.

It is best to not include trivial factors in the final analysis. This can be done by removing REACT from the original correlation matrix and refitting a $m = 4$-factor FA model. In addition, two of the 15 original variables do not correlate highly with any of the factors. These variables are DIAST and ENDUR. It may be that one or both of these variables represent, or are related to, additional factors beyond and independent of those included in the 5-factor solution. For example, we might find that the $m = 6$-factor model reproduces results very similar to the 5-factor solution but with the addition of a sixth trivial factor loading only on DIAST. It is not implausible that DIAST is independent of the five factors already identified, but it is a bit surprising that ENDUR is not correlated more highly with the third factor which we have interpretted as a measure of aerobic fitness. ENDUR is only moderately correlated with factor 1, obesity, and its low level of correlation with factor 3 suggests that factor 3 represents an aspect of cardiovascular fitness that is distinct and independent of running endurance.

To summarize, the variables in this data set seem to be determined by seven underlying characteristics. These underlying characteristics are obesity (factor 1), skeletal structure (factor 2), cardiovascular fitness (factor 3), upper body strength (factor 4), reaction time (factor 5), endurance, and diastolic blood pressure.

In the "Exercise" portion of the lab we will examine some additional analyses of the police applicant data to expand upon and to lend support to the results and interpretations obtained so far.

Exercise:

Copy the SAS program police2.sas to your home directory, run it and examine the program and its output.

In this program we fit two additional FA models to the police applicant data to try to substantiate our conjecture above that the underlying characteristics represented in this data set are obesity, skeletal structure, cardiovascular fitness, upper body strength, reaction time, endurance, and diastolic blood pressure. In the first call to PROC FACTOR we fit a 4-factor model to all of the original variables except ENDUR, DIAST and REACT. If our conjecture is correct, then we would expect that a 4-factor model will fit well, with factors representing obesity, skeletal structure, cardiovascular fitness, and upper body strength. In addition, we fit a second FA model in police2.sas based on variables ENDUR, DIAST, REACT and four variables corresponding to the factor scores for factors 1–4 (obesity, skeletal structure, cardiovascular fitness, and upper body strength). We expect to find that these seven variables are independent; that is, we expect to accept the hypothesis of no common factors in this second analysis.

1. Take a look at police2.lst. Did the ML fitting routine for the 4-factor model based on all variables except ENDUR, DIAST and REACT converge? If so, after how many iterations?

2. What is the result of the test of H_0 : 4 factors are sufficient versus H_A : more than 4 factors are needed? What is the conclusion based on this test?

3. Examine the residual correlations and partial correlations on pp.4–7 of police2.lst. Is the evidence in these matrices concerning the adequacy of the fit of the 4-factor model consistent with the LRT that I asked you about in question #2? How do the results differ?

The result observed here is not unusual. The LRT for testing whether m factors are sufficient, tends to reject the null hypothesis more often than is desirable, favoring models that include trivial factors and high values of m. According to the residual correlations and partial correlations, it appears that the 4-factor model is sufficient, which is consistent with our conjectures stated above.

4. Interpret the rotated factor loadings matrix on p.8 of police2.lst. Have we obtained 4 factors with the same interpretations as in police1.sas?

Notice that in the first call to PROC FACTOR in police2.sas I computed factor scores and requested that they be output to data set scores2. In the data step beginning DATA SCORES2; SET SCORES2;, I renamed the four new variables containing the scores on the factors obtained in the first call to PROC FACTOR. These four new variables, SKEL, OBESITY, CARDVASC, and UBSTRGTH, are then combined with the original variables REACT, DIAST, and ENDUR in the final call to PROC FACTOR.

5. Test the hypothesis that there are no common factors among these 7 variables. What is your conclusion, and how does that conclusion support or undermine our conjectures above?

If the purposes of this factor analysis included data reduction from $p = 15$ to $m < 15$ variables, the data set on p.17 of police2.lst consisting of factor scores on the 4 factors SKEL, OBESITY, CARDVASC, and UBSTRGTH, plus the observations on the three original variables REACT, DIAST, and ENDUR will be of interest. This data set can be used in any further statistical analyses that the study investigators wish to pursue.