Chapter 4: Factor Analysis

- In many studies, we may not be able to measure directly the variables of interest.
- We can merely collect data on other variables which may be related to the variables of interest.
- Goal of *factor analysis* (FA) is to relate the unobservable *latent variables* of interest to the observed *manifest variables*.
- The technique used to relate the latent variables (often called *factors*) to the manifest variables is similar to multiple regression.
- The estimation of the regression coefficients (called *loadings* in this context) is less straightforward, however.

A Factor Analysis Example: The Wechsler Adult Intelligence Study

- The Wechsler Adult Intelligence Scale (WAIS) series of tests measures participants' scores in 11 different tests.
- The multivariate data set consisted of 13 variables: these 11 test scores, plus "age" and "years of education."
- Based on the observed variables, we may want to identify certain underlying factors that cause the individuals to differ.
- Is there a "general intelligence" factor? Is there a "language ability" factor? Is there a "math ability" factor?
- Factor analysis can help us answer these questions.

Our Factor Analysis Model

- Our factor analysis model assumes that we can explain the correlations among the manifest variables through these variables' relationships with the latent variables.
- The q manifest variables are denoted $x_1, x_2, \ldots, x_q$.
- The k latent variables, or *factors*, (where $k < q$) are denoted $f_1, f_2, \ldots, f_k$.
- We relate them via a series of regression equations:

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

- The λ_{ij} values (called *loadings*) show how much each manifest variable depends on the j-th factor.
- The loading values help in the interpretation of each factor.

Our Factor Analysis Model (continued)

- We can write the regression equations in matrix notation: $\mathbf{x} = \Lambda\mathbf{f} + \mathbf{u}$, where

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{q1} & \cdots & \lambda_{qk} \end{bmatrix}$$

  and $\mathbf{f} = (f_1, \ldots, f_k)'$, $\mathbf{u} = (u_1, \ldots, u_q)'$.

- The model assumes $u_1, \ldots, u_q$ are mutually independent and are independent of the $f_1, \ldots, f_k$.

- The factors are unobserved, so we may assume they have mean 0 and variance 1, and that they are uncorrelated with each other.

Partitioning the Variance of the Data Vectors

The *communality* h_i^2 is the variability in manifest variable x_i shared with the other variables (via the factors) and ψ_i is the specific variance, not shared with the other variables.

Covariance of the Data Vectors

Hence the population covariance matrix $\mathbf{\Sigma}$ for $(x_1, x_2, \ldots, x_q)$ is $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$, where $\mathbf{\Psi} = diag(\psi_i)$.

Factor Analysis in Practice

- If this decomposition of the covariance matrix holds, then the k-factor model is correct.
- In practice, $\mathbf{\Sigma}$ is unknown and is estimated by $\mathbf{S}$ (or the sample correlation matrix $\mathbf{R}$ will be used).
- So we need to find *estimates* of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ so that the sample covariance matrix can be decomposed in this way: $\mathbf{S} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$.
- In practice, we also don't know the true value of k, the number of factors.

Methods of Estimating the Factor Analysis Model: Principal Factor Analysis

- The *Principal Factor Analysis* approach to estimation relies on estimating the communalities.
- It uses the *reduced covariance matrix* $\mathbf{S}^* = \mathbf{S} - \hat{\mathbf{\Psi}}$.
- The diagonal elements of $\mathbf{S}^*$ are $s_i^2 - \hat{\psi}_i = \hat{h}_i^2$, the (estimated) communality for the i-th variable.
- We could standardize the variables, which amounts to using the reduced correlation matrix $\mathbf{R}^* = \mathbf{R} - \hat{\mathbf{\Psi}}$.

Estimating the Communalities

- To estimate the h_i^2 values, we cannot use the factor loadings, since those have not been estimated yet.

- A more straightforward approach (when working with the correlation matrix) is one of the following:

  1. Initially let $\hat{h}_i^2$ equal the R^2 value of a regression of x_i against the other manifest variables. This is $1 - \frac{1}{r^{ii}}$, where r^{ii} is the i-th diagonal element of $\mathbf{R}^{-1}$.
  2. Initially let $\hat{h}_i^2$ equal the largest absolute correlation coefficient between x_i and any other manifest variable.

- In both of these approaches, a stronger association between x_i and the other variables will lead to a higher communality value $\hat{h}_i^2$.

- When working with the covariance matrix, we could base the communality estimates on the diagonal elements of $\mathbf{S}^{-1}$ rather than $\mathbf{R}^{-1}$.

Using the Initial Communality Estimates

- Once we have our initial $\hat{h}_i^2$ values, we can calculate $\mathbf{S}^*$ (or $\mathbf{R}^*$).
- We perform a principal components analysis on $\mathbf{S}^*$ (or $\mathbf{R}^*$) and the first k eigenvectors contain the estimates of the first k factor loadings.
- These estimated loadings $\hat{\lambda}_{ij}$ can be used to obtain new communality estimates:

$$\hat{h}_i^2 = \sum_{j=1}^{k} \hat{\lambda}_{ij}^2$$

- We can re-form $\mathbf{S}^*$ (or $\mathbf{R}^*$) with the revised communality estimates, and repeat the process until the communality estimates converge.
- This works well unless the communality estimate becomes larger than the manifest variable's total variance, implying a negative specific variance, an impossibility.

Maximum Likelihood Factor Analysis

- Maximum likelihood (ML) is a general method of estimating parameters in a statistical model.
- Classical ML requires an assumption about the form of the distribution of the data.
- If we can assume we have multivariate normal data, we can motivate a maximum likelihood estimation of our k-factor model.
- Recall that the observed sample covariance matrix is $\mathbf{S}$ and, under the factor analysis model, the true covariance matrix is $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$.
- The goodness-of-fit of the k-factor model can be judged by a "distance" measure F between the sample covariance matrix and the predicted covariance matrix under the model.

The Distance Measure and Maximum Likelihood

- Let $F = \ln|\mathbf{\Lambda}\mathbf{\Lambda}^{'} + \mathbf{\Psi}| + trace(\mathbf{S}[\mathbf{\Lambda}\mathbf{\Lambda}^{'} + \mathbf{\Psi}]^{-1}) - \ln|\mathbf{S}| - q$.
- This distance measure equals zero if $\mathbf{S} = \mathbf{\Lambda}\mathbf{\Lambda}^{'} + \mathbf{\Psi}$.
- F is large when $\mathbf{S}$ is far from $\mathbf{\Lambda}\mathbf{\Lambda}^{'} + \mathbf{\Psi}$.
- We can calculate (for a given data set) the elements of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ that make F as small as possible.
- This implies we have estimates of the communalities $h_1^2, \ldots, h_q^2$ and the specific variances $\psi_1, \ldots, \psi_q$.
- Under the assumption of multivariate normality, the likelihood $L = -0.5nF$ plus a function of the data.
- Hence minimizing F is equivalent to maximizing L.
- This method could also produce negative estimates for the specific variances.

Estimating the Number of Factors

- With factor analysis, the choice of the number of factors k is critical.
- If we use $k + 1$ factors, we will get different factors and loadings than if we use k factors.
- With too few factors, there will be too many high loadings.
- With too many factors, the loadings will be spread out too much over the factors, and the factors will be difficult to interpret.

Methods for Estimating the Number of Factors

- A subjective approach is to try various choices of k and pick the one that gives the most interpretable result — this is probably *too* subjective.
- Could use the *scree diagram* as in PCA, but the eigenvalues are not as directly interpretable in factor analysis.
- When using maximum likelihood, we can use a formal sequence of hypothesis tests to help determine k.
- We use the test statistic $U = n^{'} \min(F)$, where $n^{'} = n + 1 - (2q+5)/6 - 2k/3$.
- If the k-factor model is appropriate, this test statistic has a large-sample χ^2 distribution with degrees of freedom $(q-k)^2/2 - (q+k)/2$.
- Typically we begin with a small value of k, and increase k by 1 sequentially.
- If at any stage, the U has a non-significant P-value, we choose that value of k.
- If at any stage the degrees of freedom go to zero, the factor analysis model may be inappropriate.