

Take Home Quiz 1

Psychology 312

Spring, 2009

Note: You may use R, any R package, or LISREL to solve any aspect of this exam, or to check your work. You are also allowed to use any R routine you can find, including, of course, the library of R functions I created for your use. Points as assigned.

1. (10) Consider the following set of linear equations, with 3 unknowns, x_1, x_2, x_3 :

$$3x_1 + 4x_2 + x_3 = -3.5$$

$$x_1 - 2x_2 + x_3 = -8.5$$

$$x_1 + 2x_2 + 2x_3 = 25.5$$

a. Write this set of equations in matrix form as $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

b. Solve the set of equations for \mathbf{x}

2. (10) Given the following data in the matrix \mathbf{X} , find a 3×3 set of linear weights \mathbf{B} that produces linear transformation $\mathbf{Y} = \mathbf{XB}$ such that \mathbf{S}_{yy} , the covariance matrix of the variables in \mathbf{Y} , is exactly an identity matrix.

$$\mathbf{X} = \begin{bmatrix} 3 & 5 & 7 \\ 3 & 1 & 5 \\ 6 & 5 & 6 \\ 1 & 4 & 6 \\ 0 & 9 & 8 \end{bmatrix}$$

3. (10) Use the R help function to study the operation of the function `apply()`. Use `apply()` to create a 1-line R function `column.means()` that returns the p “column

means” of columns of data in a $N \times p$ matrix \mathbf{X} . Test it on the matrix \mathbf{X} from problem 2.

4. (10 points). In common factor analysis, we state that $\mathbf{y} = \mathbf{F}\mathbf{x} + \mathbf{e}$. Assuming that all random variables are in deviation score form, the key aspects of the model are that the common factors are standardized, that \mathbf{x} and \mathbf{e} are uncorrelated, and that $E(\mathbf{e}\mathbf{e}') = \mathbf{U}^2$, a diagonal, positive definite matrix. \mathbf{F} is the ordinary least squares regression pattern for predicting \mathbf{y} from \mathbf{x} .

If the common factors are uncorrelated and unit variance, then $E(\mathbf{x}\mathbf{x}') = \mathbf{I}$.

a. In typical common factor program output, you will sometimes see a common factor “pattern” and a common factor “structure” referred to. The common factor pattern is the pattern of regression weights for predicting \mathbf{y} from \mathbf{x} , i.e., \mathbf{F} . The common factor “structure” \mathbf{S} is the matrix of *correlations* between the observed variables in \mathbf{y} and the common factors in \mathbf{x} . Prove that, if the common factors are orthogonal, and the observed variables are in standard score form, then the common factor structure and common factor pattern are identical.

b. Suppose that the common factors are not uncorrelated but, rather, have a correlation matrix of \mathbf{P} . (Continue to assume that the factors have unit variance.) Produce a formula for the common factor structure. Assume that the observed variables are all in standard score form, so that covariances and correlations are the same.

5. (5). In common factor analysis, factors are only determined up to a nonsingular linear transformation. That is, if $\mathbf{y} = \mathbf{F}\mathbf{x} + \mathbf{e}$, then $\mathbf{y} = (\mathbf{F}\mathbf{T})(\mathbf{T}^{-1}\mathbf{x}) + \mathbf{e} = \mathbf{F}^* \mathbf{x}^* + \mathbf{e}$ for any nonsingular \mathbf{T} . (We assume that the common factors have unit variance, so there are some restrictions on \mathbf{T} .) Suppose you start with an \mathbf{x} that is orthogonal, and you rotate the factor pattern with \mathbf{T} . What will be the covariance matrix of the resulting $\mathbf{x}^* = \mathbf{T}^{-1}\mathbf{x}$?

6. (5) Briefly describe the difference between “forward” regression and “stepwise” regression in the context of multiple linear regression.

7. (25) File Regression1.sav is an SPSS file containing data for 51 variables and 500 observations. You wish to obtain an optimal linear prediction formula for predicting Y from *some* subset of the X variables. We are going to simulate a classic “cross-validation” situation where a regression equation is developed on one sample, then used to try to predict outcomes later on.

Begin by using the first 350 observations in the file as a calibration sample. You can operate on only these observations by using the SPSS *Select Cases* option on the Data menu. The remaining 150 observations are the hold-out sample, used for validation.

a. We are going to analyze the data using 4 classic approaches to regression, all of which are accessible from the SPSS menu. Use (1) standard stepwise regression, (2) forward regression, and (3) backward regression, with the default p -values both to enter and remove. Then (4) *include all variables* and do a standard regression using the “Enter” option and, to control familywise error rate, retain only variables whose regression coefficients are significant at the $.10/50 = .002$ level. What is the final model for each of the 4 methods? Adjust your output so that coefficients are printed to 6 digits of accuracy. For each model, record the *adjusted R^2* as a measure of the quality of the model.

b. Now let’s see how these models actually perform on “new data” from the hold out sample. Take the model obtained by each method, and test it on the hold-out sample (the *remaining 150 observations*) in the following way. Using the model coefficients, create $\hat{y} = \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \dots + \hat{\beta}_kx_k + \hat{\beta}_0$ for the k predictors *retained* in your model equation. Don’t forget the intercept $\hat{\beta}_0$! (Note that, in effect, this is exactly what you would do if you were using a regression equation as a prediction formula in practice.)

After computing \hat{y} , compute $e = y - \hat{y}$. Then compute the following two *cross-validation* coefficients: (a) The squared correlation between y and \hat{y} in the hold-out sample, (b) the quantity $1 - SS_e / SS_y = 1 - S_e^2 / S_y^2$.

c. Compare the *cross-validation* coefficients from part (b) with the *adjusted R^2* estimates from part (a). What do you find?

d. The formula for the adjusted R^2 is given in many texts as

$$R_{adj}^2 = 1 - (1 - R^2) \frac{N - 1}{N - k - 1}$$

where k is the number of predictors in the final model. Go back to the first model you obtained in part *a* (using stepwise regression) and verify that the adjusted R^2 printed by SPSS agrees with this formula. Show all work.

e. Several authors have suggested that, when variables are selected after viewing the data, a *better* adjustment for R^2 is to use the total number of predictors you started with, *prior to selection*, as the value of k in the adjustment formula. For all 4 of the models you obtained, compute this “conservative” adjusted R^2 . Show all work.

8. (25) An article by Long and Perkins is online. Long and Perkins reviewed a previously published measure that was never properly examined at the time of publication. The measure supposedly presented 12 items representing 4 constructs. Long and Perkins gathered data on these 12 variables, and performed two confirmatory factor analyses that they reported in their paper. One is a single factor model, and one is a 4 factor model with each factor loading on 3 variables. Here are descriptions of the 12 items:

I am going to read some things that people might say about their block. Each time I read one of these statements, please tell me if it is mostly true or mostly false about your block simply by saying "true" (2=MORE SOC) or "false" (1=LESS SOC).

SCI1 I think my block is a good place for me to live.

SCI2 People on this block do not share the same values. (reverse)

SCI3 My neighbors and I want the same things from the block.

SCI4 I can recognize most of the people who live on my block.
SCI5 I feel at home on this block.
SCI6 Very few of my neighbors know me. (reverse)
SCI7 I care about what my neighbors think of my actions.
SCI8 I have almost no influence over what this block is like. (reverse)
SCI9 If there is a problem on this block people who live here can get it solved.
SCI10 It is very important to me to live on this particular block.
SCI11 People on this block generally don't get along with each other. (reverse)
SCI12 I expect to live on this block for a long time.

NOTE: Items 2,6,8,11 are reverse coded!

Here is what I want you to do:

- a. Replicate the two (one factor and 4 factor) confirmatory factor analyses of Long and Perkins, which they report in Table 2 of their article.
- b. Using the *exploratory-confirmatory approach* of Jöreskog (1978), produce a good-fitting model with 3 factors for the same 12 SCI variables, and provide your name and description for the constructs that you think each of these three factors are capturing. Long and Perkins did a follow-up analysis in which they added several variables to the analysis and derived their own factors. How do your factors compare to theirs?