Guillaume Pagnier

April 4th, 2019

Multivariate Midterm

1. What simple but powerful form of data reduction and integration underlies any multiple

regression equation, the forming of principal component/factor scores, and contrasts in

ANOVA? (a) Name it, (b) describe what it does mathematically, (c) describe what it means

conceptually, and (d) briefly explain what role it plays in each of the three abovementioned

statistical applications (multiple regression, PCA, and contrasts).

1. Linear combinations.
2. Linear combinations represent summing up different variables (or vectors) to end up with an aggregated variable. This is done by assigning a weight (a constant) to each variable and then combining it with other variables to create a ‘weighted combination’ of the original variables. If A and B are two vectors, then 4A + 3B would be a new vector, i.e. a linear combination of A and B.
3. Conceptually, a linear combination is a type of addition/combination as you can take many different vectors or variables, assign an appropriate weight to each of them and essentially reduce those variables to a single vector.
4. In multiple regression, the regression model you’re generating to predict your DV is actually just a linear combination. For instance, the generic multiple regression model DV=B0 + B1X1 + B2X2 + … + error is a linear combination of the beta terms (the x terms represent constant weights assigned to each beta). The weights assigned are specifically designed to maximize the explained variance.

In PCA, the generated orthogonal components are just linear combinations of the original variables. The weighting constants that are chosen (in PCA’s case, it’s a weighting vector) are specifically selected to generate components that maximize the variance of the data (i.e. be able to maximally distinguish participants from one another).

Contrasts are cases of linear combinations of experimental (i.e. factor) means. An additional restriction here is that the assigned weight constants have to sum up to 0. As a result, the resulting linear combination expresses the sum of squared differences between the means of different factors and allows you to test if there is a significant difference between these factor means.

2. Reliability and validity come in several forms. (a) Describe two types of reliability and (b)

provide examples (real or invented) of conditions/situations/applications for which you would

use each. (c) Then describe two types validity and (d) provide examples of situations or

applications in which you would use each. (Note that this material was covered only partially in

the lecture itself. Refer to the additional lecture handout pages, the readings, and use other

resources, if appropriate, to answer the question.)

1. Two types of reliability would be 1) Cronbach’s alpha (describing internal consistency) and 2) interrater agreement (describing external consistency). Cronbach’s alpha represents how consistent items in a group are. This is especially used to measure scale reliability. Interrater agreement quantifies individual differences between humans when classifying subjective behavior.
2. I would use Cronbach’s alpha if I was designing a survey that gauged an individual’s extraversion using a battery of questionnaires. A high Cronbach’s alpha would indicate that my questionnaires were consistent with one another and I could trust that my battery of surveys is reliable. I would look at interrater agreement if I was scoring a video for facial expressions of fear. This is a relatively subjective analysis and there may be individual differences in my RAs in classifying what a fear expression looks like. In this case, if the scores my RAs give the same trial are similar, then I have high interrater agreement and the data would be reliable.
3. Two types of validity would be 1) content validity and 2) predictive validity. Content validity is the degree to which you’re capturing what you think you’re actually capturing i.e. is the latent construct you’re measuring the one that you’re really measuring? Predictive validity represents how valid your predictions that you’re hoping to make actually are in the real world.
4. An example of content validity would be providing an extraversion questionnaire and having that questionnaire actually be capturing extraversion, and not agreeableness or something else. In a different example, If I designed questionnaire ZZ that was meant to predict an individual’s financial success in 5 years, ZZ would have high predictive validity if it actually DID measure an individual’s success in 5 years using real world data.

3. Some people claim that exploratory data analysis (EDA) is like cheating; they argue that

looking at your data before running your significance tests biases your testing strategy and

therefore the interpretation of significance tests. Write a critical analysis of this claim, both (a)

discussing what may be correct or incorrect about it and (b) making a counterargument by

pointing to the strengths of EDA.

1. It is true that EDA could theoretically enable p-hacking. There WILL be random spurious patterns in your collected data that will appear interesting and could certainly bias the researcher towards conducting a hypothesis test they would not have conducted otherwise. In such cases, the researcher is indeed at a greater risk of making type 1 mistakes since they’re capitalizing on spurious patterns and either consciously or unconsciously driving their conclusions from these spurious results. However, this can very simply be avoided by knowing beforehand exactly by the researcher knowing what their questions are and how to answer them. If I look at some pilot data, and already know what tests I want to run and what patterns I want to look for, then it is impossible for EDA to have a negative effect since my strategy for analysis wouldn’t change when analyzing that particular set of data.
2. The strengths of EDA far, far outweigh the potential pitfall mentioned above. EDA is powerful for many reasons, one of which it TELLS you what you’re looking at. In the case of missing data or miscoded data, appropriate EDA will quickly alert you to these errors. EDA gives the researcher an idea of the distribution of the data and alerts them to any outliers or atypical responses, knowledge which is crucial when they are trying to understand what their data means. EDA allows confirms that the assumptions the researcher has in order to conduct a hypothesis test are actually true (failed assumptions – which would be invisible without EDA – can muddy up hypothesis testing). Most importantly, it allows for patterns in the data to be revealed. These patterns could be the source of future questions (to reduce the risk of a type 1 error, the researcher should replicate unexpected patterns in a new data set) or reveal different confounds that the researcher was unware of before.

4. (a) Name the type of matrix that PCA’s L matrix is. (b) Specify what all the elements (entries)

of the matrix mean conceptually. (c) Explain (verbally or formally) how SPSS, or any other

statistical program, moves from Rxx to L.

1. 𝚲 is a diagonal variance-covariance matrix of the eigenvalues. All of the off diagonals are always 0.
2. Each of the entries in the side diagonal of 𝚲 represents an eigenvalue, or the variance of a newly generated primary component. These eigenvalues are ordered such that the maximum variance is the first, the second most variance is the second and so on. All of the off diagonals, which normally represent covariance, are 0 because the generated primary components are all orthogonal to one another so there is no covariance.
3. Starting with Rxx which is a correlation matrix, SPSS conducts spectral decomposition to reconfigure R into a new ordered variance-covariance matrix, 𝚲 that consist of linear combinations of the variables of R. Spectral decomposition represents this transformation and is done by multiplying R by eigen vectors (weights) to get the new linear combinations (which end up being our generated components). How does SPSS know what eigen vectors to use? This is the crux of PCA/spectral decomposition: SPSS uses weights that will subsequently generate components that maximize the variance of the original data. Said another way, SPSS essentially creates a new component (axis) by using the covariances in R (off diagonals of R) to stretch the variance (diagonals of R) in a way that maximally differentiates between people. This would then represent one generated component. SPSS repeats the process for any variance that couldn’t be explained by the first component to create a new, second component (this is why the generated components are orthogonal to one another since all of the generated components are explaining ‘different’ variances). SPSS repeats this process until all of the variance explained in R is parsed out into the new generated components. The maximized variances of the calculated linear combinations (i.e. newly generated components) are called eigenvalues and compose the diagonal of 𝚲 in an ordered fashion with the highest number being the top number in the diagonal. As mentioned above, by nature of systematically maximizing the variance of the new components, all of the new components will be orthogonal to one another and the off diagonals of 𝚲 will be 0.

5. Write a dialogue (feel free to be funny) between a fanatic proponent of PCA and a fanatic

proponent of factor analysis. Let each person state at least *two clearly distinct* arguments for the

supremacy of their preferred technique and finish with your own (fanatic or not) conclusion.

PCA: pure transformation of variables that maximizes the variance of orthogonal linear combinations calculated from original, unaltered variables. There is only one right way to do it. Factor analysis is a model of a measurement of a latent structure. They treat the diagonal of R differently.