

9

Multivariate Analysis of Variance (MANOVA) and Discriminant Analysis

CHAPTER OBJECTIVES

- Learn how multivariate analysis of variance (MANOVA) is different from the analysis of variance (ANOVA) model.
- Why you should not necessarily perform MANOVA simply because you have several response variables.
- Understand why tests of significance in MANOVA are very different and more complex in configuration than in ANOVA, and why there is not only a single F -test as in ANOVA.
- Understand how the covariance matrix plays a fundamental role in the multivariate realm.
- Be able to compute and interpret an effect size measure for MANOVA.
- Why a multivariate model result does not necessarily break down into individual univariate findings.
- Learn what a discriminant function analysis (LDA) is, how it can be considered the “reverse” of MANOVA, and why it is sometimes used as a follow-up to MANOVA.
- Learn how to compute discriminant function scores and understand classification results based on derived discriminant functions.

In the models we have surveyed thus far, whether it is analysis of variance (ANOVA) or regression, both of these featured a **single dependent** or response variable. In both ANOVA and regression, the response variable was assumed to be a single continuous variable. Statistical models that analyze only a single dependent or response variable are generally known as **univariate models** for this reason. This is true regardless of how many independent variables are featured in the model. For example, even if an ANOVA model has three factors as independent variables, the model is still traditionally considered to be univariate since it only has a single response. Likewise, a multiple regression model may have five predictors, but is often still considered univariate since it only features a single response (though multiple regression is also often discussed under the heading of multivariate models due to its **multivariable** (i.e. several variables) nature).

Applied Univariate, Bivariate, and Multivariate Statistics Using Python: A Beginner's Guide to Advanced Data Analysis, First Edition. Daniel J. Denis.

© 2021 John Wiley & Sons, Inc. Published 2021 by John Wiley & Sons, Inc.

Thanks to developments in statistics in the early 1900s, the so-called “boon” of mathematical statistics and inference with the works of R.A. Fisher, Karl Pearson, Harold Hotelling and others, techniques were developed that allowed scientists to evaluate statistical models featuring more than a single response variable simultaneously. But why would a researcher want to analyze a model having several responses in the same model? That is a very good question and one that should be considered very carefully before deciding that a multivariate model is suitable. Before introducing and discussing these models further, one would do well to issue a few caveats on the designation of “multivariate” and briefly explore just why a researcher would want to consider a multivariate model in the first place. Many times, researchers conduct multivariate models when they would be better off with univariate ones. We address these issues next.

9.1 Why Technically Most Univariate Models are Actually Multivariate

We have defined multivariate models as those in which more than a single response variable is analyzed at the same time. However, we should note that this specification for what is multivariate vs. univariate is more of a **nomenclature** and **convention** tradition than a true distinction between univariate from multivariate at a technical level. For example, suppose we choose to model verbal ability and quantitative simultaneously as a function of training program. Perhaps some individuals were trained in one program and others a different one and we would like to see if there are **mean vector differences** (we will discuss shortly what this means) between the two training programs on both variables considered simultaneously. Though this is, by definition, a multivariate model since it has two response variables, notice that if we “flipped” the model around, such that the training program is now the response and verbal and quantitative abilities the predictors, the model could then be amenable to **logistic regression** or, as we will see, **discriminant analysis**, in which case, there is only a single response variable. Hence, what was deemed multivariate now becomes univariate very quickly, yet at their technical levels, it is a fact that these models cannot be that different from one another. In fact, they are not, as we will later briefly discuss in more detail.

The above similarities and distinctions also hold true for regression analyses. Consider two response variables as a function of a predictor. Since there is more than a single response variable, the model would rightly and formally be considered a **multivariate regression**. However, if we flipped the model around such that the single predictor is now the response, and the two responses are now predictors, the model is simply a **multiple regression**. Though the models are not identical in terms of their technical details, they are, in reality, not that far off. Hence, realize that whether you are analyzing univariate models or multivariate models, the mathematical details underlying these models are often remarkably similar, even if different elements of each model are focused on in different applications. The theoretical or research drive behind each one may be quite distinct. Given their technical similarities, **your choice of model for your research needs to be driven primarily by substantive goals.**

9.2 Should I Be Running a Multivariate Model?

There is a long tradition in statistics, especially in fields that quite often do not conduct rigorous experiments with control groups, etc., to favor methods that are statistically very complex over simpler ones. Thousands of research papers employing the latest “cutting-edge” and very complex models can be found in journal articles and other publications, so much so that it is probably a reasonable bet to assume that journal editors and reviewers will be more “impressed” if you submit a complex model than they might be with a simpler one. In this sense, it is probably the case that multivariate modeling is perceived to “advance” science in some fields much further than much simpler models. But is this true? **Does running a complex multivariate model necessarily afford a study of greater scientific merit or bring us any closer to understanding reality?** Hence, instead of starting out with why you might want to run a multivariate model, let us instead begin with why you would **not want to run such a model**. The following are the **wrong** reasons for conducting a multivariate model:

- You have several response variables at your disposal and therefore, by a “rule of thumb” that you once learned from “cookbook” teaching or reading, believe that conducting a multivariate model is the right thing to do.
- You are doubtful of running a multivariate model, but nonetheless do so because you think it will make your dissertation or scientific publication appear more “rigorous” and “advanced.”
- You believe that multivariate models are always better than univariate models regardless of the research context, since univariate models are “out of fashion” (i.e. “Nobody does ANOVA anymore, multivariate models are the ‘hip’ thing to do”).

All of the above reasons and rationales for performing a multivariate model are 100% wrong and incorrect! First off, as a general comment, only rarely should you follow any “rules of thumb” in statistics. The decision trees that fill introductory statistics and research books that tell you which analysis is correct given this or that many response and predictor variables, should be used, if at all, as simply a **guide** to choosing the most appropriate model, and nothing more. **No decision tree of this sort can tell you which model you should be running. Only you can make that decision based on your research question and knowledge of the area under investigation (and, if needed, with the advice of a statistical consultant).**

Second, a complex multivariate statistical analysis does not imply your research is more “rigorous” than if you conducted simpler analyses. **A dissertation or research publication with a poor research design and poor measurement of variables will be just as poor if you subject it to a complex statistical analysis.** Aesthetically pleasing it may be statistically, on a scientific basis it will largely be a waste of time, and, worse yet, will serve to confuse your audience regarding whether you found or did not find something in your data. Never run a complex statistical model simply to make a fashion statement, not if you want to be a serious scientist, that is. On the other hand, if your goals are social and political such that you wish to make it seem like you are advancing science through sophisticated statistics, then by all means, fit the most complex model you can find and you may just fool enough to make them think you did

something of scientific value. It is hoped, however, that **good science is your first priority**, and with this in mind, you should always choose your statistical model based on well thought out research questions, not based on technical complexity. As a scientist, your first priority is to contribute to the literature and to communicate a scientific finding, not appease or impress your audience with technical complexity. To some, you may succeed in pulling the wool over their eyes, but to others, they'll see it as it is, a **smokescreen** for a poorly run study with no experimental controls and poorly measured variables.

Finally, univariate models are never somehow “out of fashion” now that complexity rules the day in statistical modeling. For instance, in finding a cure for COVID-19, **experiments** are much more likely to use relatively simple univariate models. A 10-page paper on the cure for COVID-19 communicates more in terms of scientific utility than a 100-page paper on a poor research design flooded with multivariate and other advanced statistics. **The research design and quality of your study or experiment is always the first step to good science.** Now, from a statistical and mathematical point of view, multivariate models are indeed much more interesting and fun to work with. Our point is merely that from a scientific point of view, multivariate models do not necessarily lend more scientific credibility to your project. You are not doing pure mathematics, nor are you doing theoretical statistics. You are doing **science**, and Occam's razor and the principle of parsimony must always govern your work. **Simpler is usually better.**

The correct justification for performing a multivariate model, multivariate analysis of variance (MANOVA) in this case, is because your research question demands analyzing a linear combination of response variables at the same time. However, when might you want to do this? Let us consider a classic example, one in which a multivariate model is well-suited and appropriate. Suppose you have scores on verbal, analytical, and quant abilities on some psychometric test, such as the IQ measure we have alluded to in this book. It might make good research and theoretical sense to combine these variables together somehow, since a composite score of **verbal + analytic + quant** is probably representative in some sense of a more complete variable or construct, namely the variable you wish to analyze in this case, IQ. Hence, we can say that

$$\text{IQ} = \text{VERBAL} + \text{ANALYTIC} + \text{QUANT}$$

This is what is known as a **linear combination** of variables. A linear combination is simply a sum of variables, each weighted by a scalar, which is a number. In mathematics, the general form for a linear combination is the following:

$$l_i = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_px_p$$

where a_1 through a_p are corresponding scalars that serve as weights for variables x_1 through to x_p . In our sum for IQ, it is implied that the weights are equal to “1” on each variable. That is, our linear combination, when unpacked and spelled out in more detail, is actually the following:

$$\text{IQ} = (1)\text{VERBAL} + (1)\text{ANALYTIC} + (1)\text{QUANT}$$

Even with equal weights across all variables, the above is still defined as a linear combination, and in practice we could use it to generate the “variate” of IQ. That is, you

could, if you wanted to, sum the scores so they give you a total IQ, then use this as your response variable as a function of training program. Indeed, many times this is done in clinical research and testing on psychometric instruments without direct awareness that what is being generated is a linear combination. It is usually simply referred to as **summing the scores on the scale**, but it is not recognized that there are implied weights of “1” before each variable. So, again, we could if we really wanted to, generate the above sum and use that as our response variable. The question is, however, whether that sum can be improved upon **to satisfy a purpose other than simply being a sum of scores**. This is where MANOVA comes in. The sum that you produced with implied weights of unity before each variable is not guaranteed to have any special properties. MANOVA (and discriminant analysis) will more “wisely” select these scalars. That is, MANOVA and discriminant analysis will select these scalars so as to optimize some function of the data. As usual, and as was the case for least-squares regression, it comes down to a problem of **optimization**.

In MANOVA, the simultaneous inference occurs on a sum of variables as above, only that now the linear combination generated will not necessarily consist of scalars all (or even any) equal to 1. That is, the linear combination will no longer be chosen “naively” as we have done by simply assuming scalars of 1. Instead, the linear combination will be chosen in a way that selects the scalars with some **optimization criteria** in mind. Recall that optimization is an area of mathematics that seeks to **maximize** or **minimize** the value of a function based on particular constraints unique to the problem at hand. In other words, it selects values that are “optimal” in some sense, where “some sense” means “subject to constraints.” Techniques in calculus are often used to find such maximums or minimums. In MANOVA, optimization criteria is applied that will generate scalars for each variable, possibly all different, such that some criteria will be maximized. What is that criterion? The answer is found in the **discriminant function**. Hence, for an understanding of MANOVA, we need to first get a quick cursory glimpse of what discriminant analysis is all about, since behind the scenes of every MANOVA is a discriminant analysis. In what follows then, we quickly survey discriminant analysis before returning to our discussion of MANOVA. We will then later more thoroughly unpack the discriminant analysis model as a separate and unique statistical methodology.

9.3 The Discriminant Function

We have said that MANOVA generates a linear combination of response variables, each weighted by scalars, not necessarily all distinct, but usually so. What we need to understand now is how those scalars are selected. Using default values of “1” for each variable will not satisfy optimization requirements, but to know what scalars will, it is first important to understand which function we are trying to optimize in the first place. The key to understanding what makes MANOVA “tick” is to “flip it around” so that the multiple response variables are now predictors, and the independent variable is now the response variable. This will give us the **linear discriminant analysis model**. Recall our model thus far. We are hypothesizing IQ as a function of training program:

$$\text{IQ}(\text{verbal} + \text{analytic} + \text{quant}) \text{ as a function of Training Program (1, 2)}$$

To get the discriminant analysis, we flip this equation around:

Training Program (1, 2) as a function of IQ (verbal + analytic + quant)

In words, the **function statement** now reads that training program is a function of a linear combination of variables. In the MANOVA, we hypothesized the linear combination as a function of training program. Remarkably, these two ways of posing the question are near identical in terms of their technical underpinnings, though different details are focused on depending on which method we are studying. Hence, we are now in a position to address the question we asked previously, which is how to best select the scalars that will weight the linear sum of variables verbal + analytic + quant. Should we simply settle on scalars of “1” before each variable, or should we choose others? The answer we have come to so far is to choose scalars that will maximize some function. And it is in defining the discriminant function wherein lies the answer. **We will choose scalars that maximize the separation between groups on the binary response variable.** This linear combination, weighted by such scalars, is the same one used in the MANOVA, but we usually do not see it in our analysis. That is, it is usually not given in output or featured in the derivation of MANOVA. The discriminant analysis simply “unpacks” the linear combination so that we learn more about it. That is, we learn more about the linear combination that generated **mean vector differences** in the MANOVA. In MANOVA, mean vector differences are the focus, which, for our example, are the means on verbal, analytic, and quant as a **vector**, as a function of training program. Recall that, by definition, a vector typically has several numbers in it. For the MANOVA, the vector is one of means, which is why we call it a **mean vector**.

We will return to our discussion of discriminant analysis a bit later in the chapter. It is enough to appreciate at this point that the choice of scalars to generate the linear combination(s) in MANOVA will be determined by selecting a function or functions that maximally distinguish groups. These functions are called discriminant functions and underlie all MANOVA procedures. For now, however, we return to our discussion of MANOVA.

9.4 Multivariate Tests of Significance: Why They Are Different from the F-Ratio

We said earlier that testing multivariate models requires different tests of significance than when evaluating univariate models. The reason for this lies in the fact that in multivariate models, the simultaneous consideration of response variables requires us to model the covariance between responses. In univariate models, this covariance is not modeled since there is only a single response. When this covariance is modeled, as in multivariate models, it introduces a whole new world of complexity.

So, what does “modeling the covariance between responses” mean? To understand this, it is easiest to first introduce tests of significance for MANOVA and then unpack how each test treats this covariance. Contrary to ANOVA that featured only a single test of significance (the **F-test**), because of the potentially complex configuration of covariances in multivariate models, MANOVA features several different tests that all treat covariances a bit differently. Hence, gone are the days in which you can simply

run an F -test, obtain statistical significance, and be on your way to interpreting the accompanying effect size. In the multivariate domain, when you claim a statistically significant effect, you now need to inform your audience or reader which precise test was used that was statistically significant, as different tests may report different results. We now survey the most common multivariate tests reported by software.

9.4.1 Wilks' Lambda

The first test, and undoubtedly historically the most popular, is that of **Wilks' lambda**, given by

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}$$

We can see that the test is defined by a ratio of $|\mathbf{E}|$ to $|\mathbf{H} + \mathbf{E}|$, where $||$ in this case indicates not absolute value, but rather the **determinant** of the given matrix. The matrix in the numerator, $|\mathbf{E}|$, is a matrix of the sums of squares and cross-products that does not incorporate group differences on the independent variable. In this sense, $|\mathbf{E}|$ is analogous to **MS within** in univariate ANOVA. However, they are not one-to-one the same, since $|\mathbf{E}|$ incorporates cross-products (crudely, the aforementioned covariance idea) between responses, whereas of course MS within does not since there are no cross-products or covariance as a result of there being only a single response. \mathbf{E} is typically called the “error” matrix in MANOVA and computing the determinant on \mathbf{E} gives us an overall measure of the degree of “generalized” within-group variability in the data, again analogous in spirit to what MS within told us in ANOVA.

The matrix \mathbf{H} , like \mathbf{E} , contains sums of squares and cross-products, only that now, these sums of squares and cross-products contain between-group variation. Again, this is somewhat analogous to what is accomplished by **MS between** in ANOVA. It is often referred to as the **hypothesis matrix** for this reason, the hypothesis being that there are true population differences on the independent variable. And as is the case in ANOVA where **SS treatment + SS error = SS total**, MANOVA features an analogous identity, where $\mathbf{T} = \mathbf{H} + \mathbf{E}$. That is, in words, the total variation and cross-product variation represented by \mathbf{T} can be broken down as a sum of the hypothesis matrix \mathbf{H} and the error matrix \mathbf{E} .

So what does Λ accomplish? Notice that Wilks' is comparing, via a ratio, $|\mathbf{E}|$ to the total variation represented by $|\mathbf{H} + \mathbf{E}|$. Hence, the extent to which there is no between-group differences, $|\mathbf{E}|$ and $|\mathbf{H} + \mathbf{E}|$ will represent the same quantities, since under the case of no between-group variation, \mathbf{H} will be equal to 0, and hence $|\mathbf{H} + \mathbf{E}| = |0 + \mathbf{E}| = |\mathbf{E}|$. Therefore, we see that Wilks' is an **inverse criterion**, meaning that under the null hypothesis of no differences, $\Lambda = 1$, which is the maximum value Λ can attain. It is called an **inverse criterion** because contrary to F in univariate ANOVA, the null case is represented by a maximum value for Λ . In the univariate F -test, recall that larger values for F count as increasing evidence against the null hypothesis, and smaller values suggest that the null hypothesis is not false. For Wilks', when $\Lambda = 0$, it means that $|\mathbf{E}|$ must equal 0, and we have

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|} = \frac{|0|}{|\mathbf{H} + 0|} = \frac{0}{|\mathbf{H}|} = 0$$

Hence, the closer the value of Λ to 0, all else equal, the more evidence we accumulate against the null hypothesis. That is, the more evidence we have to reject the null hypothesis of equality of mean vectors.

9.4.2 Pillai's Trace

As mentioned, Wilks' lambda is only one of many multivariate tests. We now survey a second test, that of **Pillai's trace**, which focuses on different elements of the \mathbf{H} to \mathbf{E} components. Pillai's trace is given by

$$V^{(s)} = \text{tr}[(\mathbf{E} + \mathbf{H})^{-1}\mathbf{H}]$$

where tr is the trace of the quantity after it, that of $(\mathbf{E} + \mathbf{H})^{-1}\mathbf{H}$, and where we are taking the inverse of $\mathbf{E} + \mathbf{H}$ (recall the inverse of a matrix is denoted by $^{-1}$). But what is being assessed by $(\mathbf{E} + \mathbf{H})^{-1}\mathbf{H}$? Let us take a look at it more closely. Recall that the function of taking the **inverse** in matrix algebra is analogous (in a conceptual sense) to division in scalar algebra. Hence, the quantity $(\mathbf{E} + \mathbf{H})^{-1}\mathbf{H}$, if we were to write it in scalar algebra terms, could crudely be written as

$$\frac{\mathbf{H}}{\mathbf{E} + \mathbf{H}}$$

Thus, to understand Pillai's trace, we could say that it is comparing \mathbf{H} to the total variation represented by $\mathbf{E} + \mathbf{H}$, then taking the trace of this "quotient." We put quotient in quotes here because it will not actually be a quotient when written in proper form. In proper form it is a **product**, the product of the inverse of $\mathbf{E} + \mathbf{H}$ and \mathbf{H} . Hence, we can see that Pillai's is behaving a bit more **analogously** to traditional F in univariate ANOVA in this regard, where the hypothesis matrix is in the "numerator" (in terms of the scalar analogy) of the equation and the total variation is in the "denominator."

Pillai's can also be expressed with respect to **eigenvalues** λ_i via

$$V^{(s)} = \sum_{i=1}^s \frac{\lambda_i}{1 + \lambda_i}$$

Note that in this formulation, each respective eigenvalue λ_i is being extracted and then compared via a ratio to the denominator $1 + \lambda_i$. These ratios are then summed to provide the total value for Pillai's trace. Recall we introduced notions of **eigenvalues** and **eigenvectors** in earlier chapters. However, these elements are not intuitive and from a statistical point of view at least, they need to be grounded in a statistical application for them to have any real meaning. Hence, the interpretation of eigenvalues in the context of MANOVA will make much more sense when we survey **discriminant analysis** in more detail a bit later. Eigenvalues will make even more sense when we survey principal components later in the book as well, though those eigenvalues will not be the same eigenvalues extracted as in MANOVA or discriminant analysis.

9.4.3 Roy's Largest Root

A third multivariate test of significance is that of **Roy's largest root**, given by

$$\theta = \frac{\lambda_1}{1 + \lambda_1}$$

where λ_1 is the largest of the eigenvalues extracted. Notice that Roy's only uses the largest of the eigenvalues, whereas Pillai's uses them all. This can easily be seen by comparing the two statistics side-by-side:

$$V^{(s)} = \sum_{i=1}^s \frac{\lambda_i}{1 + \lambda_i} \quad \text{vs.} \quad \theta = \frac{\lambda_1}{1 + \lambda_1}$$

Notice that both statistics are constructed the same way. We notice, however, that in Pillai's, we are summing across the i th eigenvalues. In Roy's, we are doing no such thing, and only feature a single eigenvalue, λ_1 . **That eigenvalue is the largest of the eigenvalues extracted, and hence Roy's is focused on only the root accounting for the most information.** In the case where there is only a single eigenvalue extracted, Roy's will equal that of Pillai's, since the only eigenvalue must also simultaneously be the largest one. When would you want to use only the largest eigenvalue? The answer is when one eigenvalue dominates the size of the others, usually implying that the mean vectors lie in a single dimension (Rencher & Christensen, 2012), which, as we will see later, suggests that a **single discriminant function** is explaining most of the separation between groups. If remaining eigenvalues are sizeable and still important, then Pillai's would be preferable to compute over Roy's. All of this, again, will make more sense when we review discriminant analysis, since, as of now, it may be unclear to you why we are obtaining more than a single eigenvalue from a MANOVA problem. This cannot be adequately understood in the context of MANOVA. Discriminant analysis, however, will yield the answer.

9.4.4 Lawley-Hotelling's Trace

A fourth multivariate test is that of the **Lawley-Hotelling's trace**:

$$U^{(s)} = \text{tr}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^s \lambda_i$$

Notice that $U^{(s)}$ is taking the trace (i.e. tr) of a product, the product in this case being $\mathbf{E}^{-1}\mathbf{H}$. Recall that $\mathbf{E}^{-1}\mathbf{H}$ means we are, in analogous scalar algebra, comparing \mathbf{H} in the numerator to \mathbf{E} in the denominator. That is, we are contrasting the hypothesis matrix to that of the error matrix, where the hypothesis matrix contains the group differences and the error matrix does not. In this sense, $U^{(s)}$ is computing something somewhat similar to Pillai's discussed earlier; however in Pillai's, we took the trace of $(\mathbf{E} + \mathbf{H})^{-1}$ multiplied by \mathbf{H} . $U^{(s)}$ is also equal to the sum of eigenvalues for the problem, $\sum_{i=1}^s \lambda_i$, which is again somewhat similar in spirit to Pillai's, but does not divide by $1 + \lambda_i$ when performing the sum.

9.5 Which Multivariate Test to Use?

Though these multivariate statistics each assess something slightly different in the data as we have seen, in most cases, they will usually all hint at or strongly suggest the same decision on the multivariate null hypothesis. That is, though on a theoretical level they are different from one another (and their differences are important), in an applied sense,

you usually will not have to concern yourself about these differences as they will typically all universally suggest a similar result when applied to real data. In cases where they do not (e.g. one test suggests you reject while another does not), **you are strongly encouraged to not ignore the situation**, but instead delve into your data deeper in an exploratory sense to disentangle why one test is yielding statistical significance and another is not. As we have seen, the sizes of each statistic depend greatly on the respective sizes and patterns of **H** and **E**, and if the tests do not agree in their decision, then investigating these matrices in more detail would be called for. This would be for two purposes. The first is to better understand the results you are obtaining, but also to potentially detect patterns in these matrices that may otherwise go unnoticed. In this sense, a failure to universally reject the multivariate null across all significance tests may help you discover something in your data that is meaningful on a **scientific** as well as **statistical** level. Why did one test reject the null while another did not? The answer may (or may not!) give you insight into the nature of your data. **Never just ignore the situation. Rather, do some additional exploratory analyses to find out why.**

This is as far as we take our discussion of multivariate tests here. For further details on these tests, including comparisons and contrasts between them, the reader is encouraged to consult Olson (1976) for a classic, excellent, and still very useful overview of all of them. More theoretical multivariate statistics texts will also discuss these tests in greater detail and provide the requisite technical context for understanding and appreciating them on a deeper level. Among these include Johnson and Wichern (2007) as well as Rencher and Christensen (2012).

9.6 Performing MANOVA in Python

We now demonstrate a MANOVA in Python. For this demonstration, we consider the aforementioned data, but only on quant, verbal, and train group. We first create the small data file:

```
import pandas as pd
data = {'quant': [5, 2, 6, 9, 8, 7, 9, 10, 10],
        'verbal': [2, 1, 3, 7, 9, 8, 8, 10, 9],
        'train': [1, 1, 1, 2, 2, 2, 3, 3, 3]}

df_manova = pd.DataFrame(data)
print(df_manova)
   quant  verbal  train
0      5      2      1
1      2      1      1
2      6      3      1
3      9      7      2
4      8      9      2
5      7      8      3
6      9      8      3
7     10     10      3
8     10      9      3
```

We confirm in this data that there are a total of **eight** cases, with three cases per train group. We now run the MANOVA using **statsmodels**. Before we run the MANOVA, we need to first label **train** as a **categorical variable** so it will be recognized in the model statement. As it stands, it is a **numeric variable**, as we can easily confirm:

```
print(df_manova.dtypes)
quant      int64
verbal     int64
train      int64
dtype: object
```

Indeed, we see from the above that all variables are being identified as numeric. For **quant** and **verbal** this is how we want them, as they are continuous variables in the MANOVA. However, for **train**, we want this to be a factor variable, so we convert it to a categorical one:

```
cols = ['train']
for col in cols:
    df_manova[col] = df_manova[col].astype('category')
print(df_manova.dtypes)
quant      int64
verbal     int64
train      category
dtype: object
```

We confirm above that **train** is now a categorical variable and we can now proceed with the MANOVA:

```
from statsmodels.multivariate.manova import MANOVA
maov = MANOVA.from_formula('quant + verbal ~ train', data = df_
manova)
print(maov.mv_test())
```

```

                Multivariate linear model
=====
-----
                Intercept      Value  Num DF  Den DF  F Value  Pr > F
-----
                Wilks' lambda  0.1635  2.0000  5.0000  12.7885  0.0108
                Pillai's trace  0.8365  2.0000  5.0000  12.7885  0.0108
                Hotelling-Lawley trace  5.1154  2.0000  5.0000  12.7885  0.0108
                Roy's greatest root  5.1154  2.0000  5.0000  12.7885  0.0108
-----
                train          Value  Num DF  Den DF  F Value  Pr > F
-----
                Wilks' lambda  0.0561  4.0000  10.0000  8.0555  0.0036
                Pillai's trace  1.0737  4.0000  12.0000  3.4775  0.0417
                Hotelling-Lawley trace  14.5128  4.0000  5.1429  17.8112  0.0033
                Roy's greatest root  14.3516  2.0000  6.0000  43.0547  0.0003
=====
```

The output for the intercept is of no use and we do not interpret it. We are more interested in the output for `train`. We can see that all tests of significance, from Wilks' lambda through to Roy's largest root, are statistically significant at least at $p < 0.05$, but what was our null hypothesis to begin with? Recall that the MANOVA analysis is not being done on means. Rather, it is being performed on **mean vectors**. That is, the null hypothesis for this problem is that mean vectors are equal across train groups:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$$

where the boldface type $\boldsymbol{\mu}_1$ and so on are now **vectors**, not univariate means as in ANOVA. The alternative hypothesis is that somewhere amid these mean vectors is at least one pairwise difference. For our data then, we reject the null hypothesis of equality of mean vectors and conclude that at least one mean population vector on the `train` variable is different from another.

9.7 Effect Size for MANOVA

Having computed the significance tests for MANOVA and rejected the multivariate null, we would now like to compute an **effect size**. Recall the reason for computing effect sizes. Effect sizes inform us of the actual variance explained in our data and are not as influenced by sample size as are p -values. An effect size for Wilks' lambda can be computed as

$$\eta_p^2 = 1 - \Lambda^{1/s}$$

where η_p^2 is **partial eta-squared**, Λ is the Wilks' value from the MANOVA, and s is the smaller of either the number of dependent variables or the degrees of freedom for the independent variable. Since for our data $\Lambda = 0.056095$ and s is equal to 2 since the minimum of 2 dependent variables and 2 degrees of freedom is equal to 2 (we will discuss shortly in discriminant analysis where this rule comes from), the computation we need is the following:

$$\eta_p^2 = 1 - \Lambda^{1/s} = 1 - 0.056095^{1/2} = 1 - 0.23684383 = 0.76$$

Hence, the effect size associated with our multivariate result is equal to 0.76. Recall that in this case we are analyzing a linear combination and not individual variables. Hence, we are reporting the variance explained across a linear combination of variables, not a single variable as in ANOVA. There are some complexities that arise when using η_p^2 in a MANOVA context. For details, see Tabachnick and Fidell (2001, p. 339). This is as far as we discuss effect sizes in MANOVA. For a more thorough discussion of effect sizes in multivariate settings in general, see Rencher and Christensen (2012).

9.8 Linear Discriminant Function Analysis

Having surveyed the MANOVA model, we now consider in more detail **linear discriminant function analysis**. As mentioned earlier, on a technical level LDA ("linear discriminant analysis") can be considered as the "reverse" of MANOVA. That is, if you

find a statistically significant effect in MANOVA, it suggests mean vector differences on a linear composite of response variables. The question asked by discriminant analysis is:

What is the nature of this one (or more) linear composite(s) that is (are) successful in providing group differentiation?

Here is a great way to think of what a discriminant function actually is, by way of analogy. Suppose you tell your friends that you are a master of being able to determine who is intelligent vs. who is not, simply by talking and watching them for a few minutes. So, you meet someone, and after an hour of talking with them, you draw the conclusion that they either belong in the intelligent group or the group not-so-intelligent. Suppose out of 100 trials with each time a different subject, you are correct every time. That is, you are able to correctly classify subjects into the correct IQ group. Your friends, of course, would ask you:

How are you so good at discriminating intelligent vs. non-intelligent people?

In other words, what they are asking is the nature of the **discriminant function** you are using to successfully discriminate intelligent vs. not-so-intelligent people. What might your discriminant function be made of? Let us consider some possibilities that you may be using, on an informal level, to determine group membership:

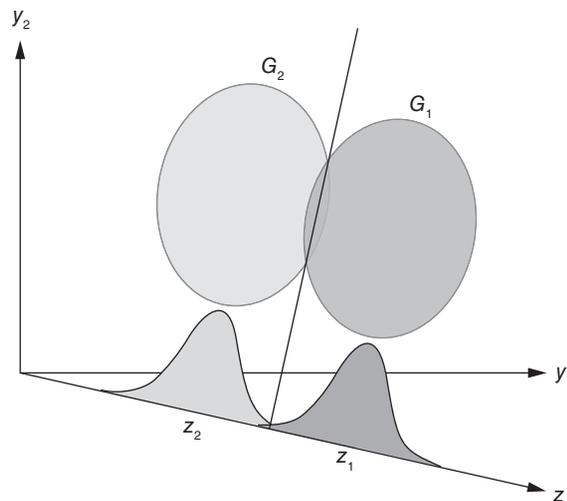
- **Quickness of speech** of the subject, in that people who speak faster seem to be “sharper,” and so if they are speaking quickly and seem sharp, this variable contributes to them being classified into the “intelligent” group.
- **Use of vocabulary** by the subject; maybe you believe subjects who use a bit more advanced vocabulary belong in the intelligent group.
- **Degree of delay in responding in communication** by the subject, in that after you are finished speaking, the elapsed time until they start speaking might indicate their degree and speed of processing the information.

Of course, we are making up the aforementioned criteria for the convenience of our example. However, the point is that in your discriminating ability are likely factors or variables you consider mentally such that they help you group people into one category or another. This is your discriminant function for grouping individuals, and might be made up, informally, of the above variables, as one example.

The question then becomes, how much respective “weight” do you assign to each of the above variables? That is, do you deem quickness of speech much more “important” than use of vocabulary and delay in responding? If you do, then perhaps intuitively you are allowing that variable to carry more **impact** or **influence** in your discriminant function. In discriminant analysis, we will no longer rely on your “informal criteria,” but rather will attempt to arrive at a discriminant function via much more precise and intelligent means. In other words, we will develop some **optimization criteria** for deriving the **best** discriminant function. This is the essence of what discriminant analysis is about, to obtain **scalars** (weights) such that when applied to the variables, best serve to discriminate between groups. That is, the selection of scalars maximizes some criterion, in a similar way that least-squares regression minimized some criterion (i.e. the sum of squared errors in prediction).

9.9 How Many Discriminant Functions Does One Require?

The question of how many discriminant functions should be kept from an analysis is different from the question of how many discriminant functions are generated. For a given problem, the number of discriminant functions generated will be equal to the **smaller of the number of predictor variables or one less the number of groups on the dependent variable**. Hence, if there are p populations on the dependent variable, then in general there will be $p - 1$ discriminant functions generated. At first glance, you may think the number should equal p rather than $p - 1$, but this intuition would be wrong. Why is this so? We can demonstrate why through our previous example of the mental discriminant function we used to distinguish among intelligent folks vs. not-so-intelligent people. Recall that your criteria consisted of three variables to do the discriminating, but there are only two groups to discriminate. Hence, you can think of those three variables as representing a new “variate,” a new axis upon which the discriminating will take place. Now, how many variates will you need to discriminate between two groups? Only a single function will do the job, because that single dimension can “cut through” the two categories to provide separation, as shown in the following diagram:



Notice that in the figure, the newly derived function is the line that is providing maximum separation between the groups **A** and **B**. Only a single discriminant function is needed to separate two groups. In our analogy, you only need one **mental discriminator** (in this case) to distinguish between intelligent vs. not-so-intelligent people. Now, for three groups, the maximum number of functions is two, and what is more, one of those functions may do a better job at discriminating than the other. Note as well the overlap in the diagram. There is overlap because the function is not doing a perfect job. If the function is not doing a perfect job, then why use it? It is key to understand that while the discriminant function will rarely provide perfect separation between groups, it is not advertised to do so in the first place. Rather, it is designed to provide **maximum separation** between groups. This is key. But what does “maximum separation” mean?

At first glance, it would seem to suggest “perfect” separation, but this is not the case, no more than a least-squares line promises to fit data perfectly or provide perfect prediction. The least-squares line is simply designed to minimize the sum of squared errors, or, equivalently, maximize R-squared (both amount to the same thing). It promises to do this better than any other line one could fit to the data, and so it is with the discriminant function. The discriminant function provides maximum separation even if the separation is far from perfect. What will determine the degree of separation achieved? The quality of the data and the true existence of mutually exclusive groups will determine how good the separation is. Beyond that, the discriminant function can only do the best it can, just as any other statistical method is only as good as the data on which it is fit. Again, as we have highlighted throughout this book, **statistical methods do not “save” data, they simply model it.** The extent to which a model fits well is up to your science, not the model. All the statistical method can guarantee is that it is doing the best it can. That is, all the statistical method or algorithm can do, in a very true sense, is usually to maximize or minimize a function. Despite what researchers want to believe of their statistics, statistical modeling usually cannot do much more than that. The rest is up to you and your research design.

9.10 Discriminant Analysis in Python: Binary Response

We demonstrate a couple of simple examples of LDA in Python. Our first example features two predictors on a binary response variable, that is, a response variable having only two categories. After this, we will consider the example of three levels on the response variable, that is, discriminant analysis on a **polytomous** response. We first build our data frame for the two-case:

```
data_discrim = {'y': [0, 0, 0, 0, 0, 1, 1, 1, 1, 1],
                'x1': [4, 3, 3, 2, 2, 8, 7, 5, 3, 3],
                'x2': [2, 1, 2, 2, 5, 3, 4, 5, 4, 2]}
```

```
df_discrim = pd.DataFrame(data_discrim)
```

```
df_discrim
```

```
Out[47]:
```

	y	x1	x2
0	0	4	2
1	0	3	1
2	0	3	2
3	0	2	2
4	0	2	5
5	1	8	3
6	1	7	4
7	1	5	5
8	1	3	4
9	1	3	2

For these data, we would like to discriminate groups 0 and 1 on y using predictors x_1 and x_2 . This will call for a single discriminant function to be produced since there

are only two categories on the response. Recall that even had we many more predictors, since there are only two response categories, we can only produce a single discriminant function. As an analogy, if I asked you to guess which hand behind my back contains the quarter, you require only a single “mental function” to make your choice; it is either in the right hand or the left hand, and hence whatever information (i.e. “linear combination” by analogy) you use to make this discrimination can be combined into a single function, in this case, a single linear combination. You only use one discriminating tool to make the decision, even if that tool is rather complex and made up of a linear sum of different variables.

Let us build our response and predictor sides. We will use `np.array()` for this:

```
y = np.array([0, 0, 0, 0, 0, 1, 1, 1, 1, 1])
X = np.array([[4, 2], [3, 1], [3, 2], [2, 2], [2, 5], [8, 3], [7, 4],
             [5, 5], [3, 4], [3, 2]])
```

To demonstrate that Python will be unable to generate more than a single function, suppose we requested two functions via `n_components=2`:

```
import numpy as np
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
lda = LinearDiscriminantAnalysis(n_components=2)
lda.fit(X, y)
```

```
FutureWarning: In version 0.23, setting n_components > min(n_features,
n_classes - 1) will raise a ValueError. You should set n_components to None
(default), or a value smaller or equal to min(n_features, n_classes - 1).
```

```
warnings.warn(future_msg, FutureWarning)
```

The reason we received an error is that we should be requesting at most a single function (for reasons just discussed), as in the following correct implementation of the code where we specify `n_components=1`:

```
lda = LinearDiscriminantAnalysis(n_components=1)
model = lda.fit(X, y)
```

```
Out[257]:
```

```
LinearDiscriminantAnalysis(n_components=1, priors=None,
                           shrinkage=None,
                           solver='svd', store_covariance=False,
                           tol=0.0001)
```

The following are the discriminant function scores for the function:

```
scores = lda.transform(X)
scores
```

```
Out[259]:
```

```
array([[-0.43107609],
       [-1.35954768],
       [-0.92847159],
       [-1.42586708],
```

```

[-0.1326388 ],
[ 1.98958197],
[ 1.92326257],
[ 1.35954768],
[-0.0663194 ],
[-0.92847159]])

```

That is, the first case in our data via the model, obtained a discriminant score of -0.43 , the second case -1.36 (rounded up), and so on. We can obtain the coefficients to the discriminant function using `lda.coef_`:

```

print(lda.scalings_)
[[0.49739549]
 [0.43107609]]

```

These are the raw **unstandardized** discriminant function coefficients. The constant for the function is equal to -3.283 , and hence we can now use the function to generate discriminant scores:

$$y = -3.283 + 0.49739549(x_1) + 0.43107609(x_2)$$

For an example regarding how to use the function, our first observation in our data has a score of 4 on x_1 and 2 on x_2 . Let us compute that case's discriminant score:

$$\begin{aligned}
 y &= -3.283 + 0.49739549(4) + 0.43107609(2) \\
 y &= -3.283 + 1.98958196 + 0.8621521 \\
 &= -0.43126586
 \end{aligned}$$

Notice that the computation agrees (within rounding error) with the first score generated by Python earlier. We can duplicate all of the scores that Python produced automatically using `lda.transform(X)` earlier (output not shown).

The value of each discriminant function evaluated at group means can be computed:

```

m = np.dot(lda.means_ - lda.xbar_, lda.scalings_)
m

Out[265]:
array([[ -0.85552025],
       [ 0.85552025]])

```

That is, the mean of the function at $y = 0$ is equal to -0.855 while the mean of the function at $y = 1$ is 0.856 . Next, we will use the discriminant functions to predict group membership using `model.predict()`:

```

pred=model.predict(X)
pred
Out[195]: array([0, 0, 0, 0, 0, 1, 1, 1, 0, 0])

```

We can see that the model correctly predicts the first five cases (0, 0, 0, 0, 0), which recall were all in the designated "0" group. It also correctly classifies the first three cases of the five cases in group "1" (1, 1, 1, 0, 0). Notice it misclassifies the last two cases

(0, 0). Hence, cases 9 and 10 were misclassified. We can obtain a convenient summary of this via what is known as a **confusion matrix**:

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score

print(confusion_matrix(pred, y))

[[5 2]
 [0 3]]
```

Notice that if all classifications were correct, the “5” and “3” across the main diagonal would equal “5” and “5.” The “2” in the upper right of the matrix corresponds to the two misclassified cases.

9.11 Another Example of Discriminant Analysis: Polytomous Classification

We demonstrate a second discriminant analysis, this time on the `train` data. First, we set up our data file:

```
train = np.array([1, 1, 1, 2, 2, 2, 3, 3, 3])
X = np.array([[5, 2], [2, 1], [6, 3], [9, 7], [8, 9], [7, 8], [9, 8],
             [10, 10], [10, 9]])
```

X

Out [65]:

```
array([[ 5,  2],
       [ 2,  1],
       [ 6,  3],
       [ 9,  7],
       [ 8,  9],
       [ 7,  8],
       [ 9,  8],
       [10, 10],
       [10,  9]])
```

```
model = lda.fit(X, train)
```

model

Out [68]:

```
LinearDiscriminantAnalysis(n_components=None, priors=None,
                           shrinkage=None,
                           solver='svd', store_covariance=False,
                           tol=0.0001)
```

Since we have three groups on the response variable, Python will compute for us two functions. As before, to obtain the raw coefficients (which, incidentally, agree with SPSS’s unstandardized coefficients if you also conduct the analysis using that software), we compute:

```
print(lda.scalings_)

[[ 0.02983363  0.83151527]
 [ 0.9794679 -0.59019908]]
```

As we did for the two-group case, we can obtain the discriminant scores for each function (which also agrees with SPSS):

```
lda.transform(X).shape
Out[75]: (9, 2)

lda.transform(X)

Out[76]:
array([[ -4.31397269,  0.61732703],
       [-5.38294147, -1.28701971],
       [-3.30467116,  0.85864322],
       [ 0.70270131,  0.99239274],
       [ 2.63180348, -1.01952069],
       [ 1.62250195, -1.26083688],
       [ 1.68216921,  0.40219366],
       [ 3.67093863,  0.05331078],
       [ 2.69147074,  0.64350986]])
```

Finally, we can obtain the model prediction classifications (which, again, agrees with SPSS's discrim function):

```
pred = model.predict(X)
pred
Out[114]: array([1, 1, 1, 2, 2, 2, 2, 3, 3])
```

We see that only a single case was misclassified. This is reflected in the confusion matrix for the classification:

```
print(confusion_matrix(pred, train))
[[3 0 0]
 [0 3 1]
 [0 0 2]]
```

9.12 Bird's Eye View of MANOVA, ANOVA, Discriminant Analysis, and Regression: A Partial Conceptual Unification

In this chapter, we have only had space to basically skim the surface of all that makes up MANOVA and discriminant analysis. A deeper study will emphasize eigenvalues much more, the dimension reduction that takes place in each procedure, as well as the assumptions underlying each analysis, such as equality of population covariance matrices on levels of the categorical grouping variable. Entire chapters and books are written on these procedures, and hence beyond our basic cursory overview and simple demonstration in Python, additional sources should be consulted. Denis (2021)

provides a deeper and richer overview, and Rencher and Christensen (2012) survey these techniques in much greater detail than in this chapter. Navigating the multivariate landscape can be challenging, and hence we close this chapter with an overview of how many models can be subsumed under multivariate models in general.

MANOVA, for whatever reason, has historically not received the “fame” that discriminant analysis has received, even though, as we have seen, it is essentially the “reverse” of discriminant analysis. Today, with **data science** and **machine learning** being quite popular, MANOVA seems to have taken a backseat to discriminant analysis, largely due to, I believe, the **substantive focus** of many of the problems in those areas of investigation. In machine learning, for instance, the focus is on **classification**, not on **mean vectors**, and less on the nature of the functions generated by the discriminant analysis. The emphasis is less on **theoretical construction** and **justification of dimensions** as it is on simply getting the classification “right” and optimizing prediction accuracy. “Black boxes” are more “tolerable” in these fields than in fields where identifying and naming dimensions is a priority. So long as you can predict the correct category, it does not necessarily matter as much what “ingredients” went into this prediction. This is what we mean by the “black box.” If it works, it works, and that is what matters. At least this is the primary emphasis in some fields of quantitative practice.

In traditional statistical applied substantive practice, however, for good or for bad, scientists would nonetheless like to often “name” the dimensions emanating from the discriminant function analysis, in addition to learning how well they can predict and classify. **The theoretical nature of the discriminator is of great interest.** Hence, in many applied fields, knowing that one or two dimensions can successfully predict and classify is not enough; researchers would like to attempt to **identify such dimensions**, which is done by focusing on the coefficients (usually the standardized ones) resulting from the analysis. Coefficients that are greater in absolute value typically indicate that the given variable is more relevant or “important” (in some sense) to the discriminant function than coefficients with lesser absolute value.

As it pertains to mean vector differences in MANOVA, this also does not seem to be a priority of many in data science or machine learning as it was in the day when Wilks introduced lambda in 1932. However, for the applied researcher, the theory of a linear combination of variables as a function of one or more independent variables is still just as relevant today and many questions in the social and natural sciences require thinking of problems in exactly this way. The concept of “latent variables” is common in many of these sciences, and hence a focus on eigenvalues, multivariate significance tests, and all the rest of it becomes quite relevant when conducting such analyses. ANOVA and MANOVA are not somehow “out of vogue” or “outdated” methodologies. From a pedagogical point of view, understanding these methods paves the way to understanding relatively advanced statistical techniques such as multilevel modeling and certain very advanced psychometric methodologies.

The “gist” of the above comments is that simply because we are approaching a problem in one way or another does not imply a lack of underlying similarity among the approaches. This is precisely the case as it concerns statistical modeling. While each statistical approach has its own peculiarities, it is also true that they are more **unified** than they are **disparate**. And actually, we can identify one statistical method for which many other models can be considered “special cases.” What is that model?

Canonical correlation. A **canonical correlation** model has several variables on each side of the equation,

$$y_1, y_2, y_3, \dots, y_j = x_1, x_2, x_3, \dots, x_k$$

where on the left-hand side are variables y_1 through y_j , while on the right-hand side are variables x_1 through x_k . All of these variables are assumed to be continuous. But what does a canonical correlation measure? **Canonical correlation assesses the linear relationship among linear combinations of y variables to x variables**, where the y variables and x variables, not surprisingly, are each weighted by **scalars** to maximize the bivariate correlation between linear combinations:

$$(a_1)y_1 + (a_2)y_2 + (a_3)y_3 + \dots + (a_n)y_j = (b_1)x_1 + (b_2)x_2 + (b_3)x_3 + \dots + (b_n)x_k$$

where a_1 through a_n are corresponding scalars for the left-side linear combination and b_1 through b_n are corresponding scalars for the right-side linear combination (usually different from those on the left-hand side). In canonical correlation, as with Pearson r , there are no dependent or independent variables. There are simply two sides of the equation as one would have in a bivariate Pearson r correlation. The scalars a_1 through a_n and b_1 through b_n are chosen to **maximize the bivariate correlation between linear combinations** in the same spirit that predicted values in regression were obtained in order to maximize the bivariate correlation between predicted and observed values in simple or multiple linear regressions. Canonical correlation underlies most of the statistical methods we have surveyed to date, even if output for each procedure does not report it. **Canonical correlation is a fundamental idea of multivariate analysis, which is to associate linear combinations between sets of variables (left side and right side).**

9.13 Models “Subsumed” Under the Canonical Correlation Framework

Now, consider how virtually all (well, many at least) of the models we have talked about thus far can be considered “special cases” of the wider canonical correlation framework, simply by **changing how variables are operationalized** for each given model. If we want a **multivariate multiple regression**, then we use the right-hand side to predict the left-hand side. Hence, we are regressing the linear combination $y_1, y_2, y_3, \dots, y_j$ onto $x_1, x_2, x_3, \dots, x_k$. The research question is to learn whether the linear combination $(a_1)y_1 + (a_2)y_2 + (a_3)y_3 + \dots + (a_n)y_j$ can be predicted by the linear combination $(b_1)x_1 + (b_2)x_2 + (b_3)x_3 + \dots + (b_n)x_k$. The scalars of each linear combination a_1 through a_n and b_1 through b_n (usually distinct) are chosen such as to maximize this predictive relationship, analogous to how scalars in multiple regression in the regression equation, $y_i = a + b_1x_{1i} + b_2x_{2i} + \dots + b_px_{pi} + \varepsilon_i$, are chosen to maximize the correlation between “sides” of the multiple regression equation. The only difference is that in multivariate multiple regression, each side consists of a linear combination instead of as in multiple regression where the right-hand side is a linear combination and the left-hand side is simply y_i . Or, if you wish to make things even more general (i.e. a bird flying really high!), you can think of the left side y_i as a very **simple linear combination made up of only a single variable**. Even if the analogy is not perfect,

the most general conceptual framework pays dividends when it comes to trying to place these models in context. That is, single variables can be conceived as very simple and elementary linear combinations where all other possible variables are scaled with values of zero. Multivariate analysis typically analyzes more complex linear combinations. When we were doing a simple *t*-test, for instance, we could, in this framework, conclude the response variable to be a very basic linear combination consisting of only a single variable.

What about good 'ol **ANOVA**? Recall in ANOVA that the left-hand side is a single continuous response variable, while the right-hand side is one (one-way ANOVA) or more (factorial ANOVA) independent variables, each scaled as a categorical factor variable. How can this be obtained from the wider canonical correlation model? Simply reduce the left-hand side to a single continuous variable and change the right-hand side to a categorical variable(s) with factor levels:

$$y_1, y_2, y_3, \dots, y_j = x_1, x_2, x_3, \dots, x_k \quad \text{Canonical model}$$

$$y_1 = x_1(1,2), x_2(1,2,3), x_3(1,2,3,4), \dots, x_k(1, \dots, n) \quad \text{ANOVA (one-way through factorial)}$$

where y_1 is the single continuous response, $x_1(1,2)$ is a categorical factor with two levels, $x_2(1,2,3)$ is a categorical variable with three levels, and so on, to represent factorial ANOVA. Hence, we see that **the ANOVA model can be conceptualized as a special case of the wider canonical model**. Formally, it will be different of course, but the above conceptualization is a powerful way to understand how ANOVA relates to the "bigger" more inclusive canonical model. All we need to do is change the coding on the right-hand side of the equation and simplify our "linear combination" on the left-hand side.

How about **MANOVA**? Add more response variables, while keeping the right-hand side of the equation the same as in the ANOVA:

$$y_1, y_2, \dots, y_j = x_1(1,2), x_2(1,2,3), x_3(1,2,3,4), \dots, x_k(1, \dots, n)$$

How about **discriminant analysis**? Flip the MANOVA model around, and for binary discriminant analysis, reduce the number of variables to one, with two levels:

$$y_1, y_2, \dots, y_j = x_1(1,2), x_2(1,2,3), x_3(1,2,3,4), \dots, x_k(1, \dots, n) \quad \text{MANOVA}$$

$$x_1(1,2) = y_1, y_2, \dots, y_j \quad \text{DISCRIM}$$

Hence, we see that several of the models we have surveyed in this book can be conceptually (and to a remarkable extent, technically as well) subsumed under the wider canonical correlation model. In a deeper study of these relationships, it would be acknowledged that an **eigenvalue** in one-way ANOVA can be extracted, which will be equal to **SS between** to **SS error**, which is the exact same eigenvalue that is extracted when performing the corresponding discriminant analysis, where the response is the factor of the ANOVA and the predictor is the continuous response. That is, the **eigen analysis**, to a great extent, subsumes both analyses. There is good reason why earlier in the book it was said that eigen analysis is the underground framework to much of statistical models. Eigen analysis is, in a strong sense, what univariate and multivariate statistics are all about, at least at the most primitive technical level. Much of the rest is in the details of how we operationalize each model and the purpose for which we use it. Always strive to see what is in the deepest underground layer of statistical

modeling and you will notice that it is usually a remarkably simple concept at work that unites. From “complex” models to sweet simplicity when you notice common denominators. **In statistics, science, and research, always seek common denominators, and you will learn that many things are quite similar or at minimum have similar “tones” in meaning.**

Music theory, for instance, at its core is quite simple, though you would never know it by how the piano player makes it sound complex! But when you understand what the piano player is doing and the common denominator to most jazz arrangements, for instance, you are better able to appreciate the underlying commonality. Unfortunately, you often need to experience a certain degree of complexity before you can “see” or appreciate the simplicity through the complexity. Likewise, if I told you from the start of the book that many statistical models can in a sense be considered special cases of canonical correlation, it would have made no sense. It probably makes more sense now because you have surveyed sufficient complexity earlier in the book to be in a position to appreciate the statement and “situate” the complexity into its proper context.

Hence, though statistical procedures are definitely not completely analogous to one another and do have important technical differences, we nonetheless can appreciate that different procedures have a similar underlying base. That is, the machinery that subsumes many “different” statistical methods is often quite similar at some level, and understanding a bit of this “underground” immediately **demystifies the different names we give to different statistical methodologies**. This is as far as we take these relations here in this book as the deeper comparisons are well beyond the scope of what we can do here, but if you study and explore these relations further, you will no longer perceive different names for many statistical methods as necessarily different “things” in entirety, thereby falling prey to semantics and linguistic distinctions. You will see them as special cases of a more **unified and general framework**, and a certain degree of “statistical maturity” will have been attained in consequence (Tatsuoka, 1971).

Review Exercises

1. What is the difference between a **univariate** and a **multivariate** statistical model? Explain.
2. Why are most models actually **multivariate** from a technical point of view even if they are not typically named as such?
3. Why is using a “**decision tree**” alone potentially unwise when selecting which statistical model to run? Why is it important to merge the **scientific objective** with the choice of statistical model?
4. How can running a **complex multivariate model** be virtually **useless** from a scientific point of view?
5. What is meant by a **linear combination** of variables, and how does this linear combination figure in the MANOVA?
6. In the example featured in the chapter, why is the **linear combination IQ = (1) VERBAL + (1)ANALYTIC + (1)QUANT** considered a “naïve” one? Why is this particular linear combination likely to not be the one chosen by MANOVA?

7. What is **discriminant analysis** and how is it related to **MANOVA**?
8. Why are **multivariate** tests of statistical significance necessarily more complex than **univariate** tests? What do multivariate tests incorporate that univariate tests do not?
9. Prove that **Wilks' lambda** has a minimum value of 0 and a maximum value of 1.0.
10. How is the **trace** involved in Pillai's trace? That is, why take the trace at all? Why is multiplying the inverse of $\mathbf{E} + \mathbf{H}$ to \mathbf{H} not enough? Unpack Pillai's trace somewhat and explain.
11. When would interpreting **Roy's largest root** be most appropriate? Why?
12. Consider the **iris** data and run a **MANOVA** on a linear combination of iris features as a function of species. Interpret Wilk's lambda, and compute an effect size for the MANOVA. Is the result statistically significant, and if so, how much variance is accounted for?
13. On the same **iris** data as in Exercise 12, perform now a **discriminant function analysis** where species is the response variable and the set of features is the predictor space (select two species of your choice). Relate the MANOVA to the discriminant analysis results in as many ways as you can.
14. Conceptually justify the **number of discriminant functions** computable on a set of data. For example, if there are three levels on the response, why is the number of functions equal to 2? If there are two levels on the response, why is the number of functions equal to 1? Explain.