4

Regression Mixtures

This primer focuses on regression mixture modeling. We assume you have read the section on it in Chapter 11, but repeat parts of it here to set context. We also assume you are familiar with multiple regression.

The traditional regression model expresses an outcome, Y, to be a linear function of predictors (in this case, X1 and X2) in accord with a linear equation:

$$Y = \alpha + \beta_1 X1 + \beta_2 X2 + \varepsilon$$

where α is the intercept, the β are regression coefficients, and ε is an error term. When we conduct a regression analysis, we implicitly assume the individuals represent a single population with common parameter values in the equation. It is possible, however, that the population is composed of mixtures of sub-populations that have different parameter values, namely different regression coefficients and/or different intercepts. In essence, the population is a mixture of two or more sub-populations (also called *segments*) with one or more distinct regression coefficients or intercepts characterizing the relationship between Y and the various X. In such a scenario, the population regression model can be misleading and mischaracterize the separate population segments. If the source of such population mixing is known, then it can be modeled through interaction terms in the regression equation. In cases where the source of mixing is unknown, one can seek to identify the population segments on an exploratory basis using *regression mixture modeling* (Vermunt & Magidson, 2004; Muthén & Asparouhov, 2009). Regression mixture modeling combines conventional regression models with methods known as latent class models (Lazarsfeld & Henry, 1968; McCutcheon, 1987) in an attempt to identify unknown population mixtures empirically. Regression mixture modeling is particularly useful when the sources of coefficient heterogeneity have not been thoroughly thought out *a priori*. Regression mixtures can provide evidence for the existence of subgroups that then can be further studied in future research. The central task of regression mixture analysis is to identify the number of heterogeneous segments there are and classify individuals into the different segments. The task of the analyst then becomes to give substantive meaning to the identified segments.

In this primer, we first describe the mechanics and logic of regression mixture modeling. Next, we consider how one decides on the number of heterogeneous segments

that exist in a population relative to the regression model being explored. This requires providing background on statistics derived from information theory, namely the Akaike Information Criterion and the Bayesian Information Criterion. After providing an intuitive sense of these statistics, we discuss confusion matrices and their role in identifying the number of heterogeneous segments in a population. Finally, we consider the issue of giving the segments substantive meaning and interpretation.

REGRESSION MIXTURES: THE MECHANICS

Regression mixture models use a categorical latent variable to describe the means and covariances of observed data (Magidson & Vermunt, 2004), with the latent variable defining a mixing of subpopulations each with distinct multivariate distributions of the observed variables. The latent variable is often called a *latent class variable*. In regression mixture modeling, the latent class variable is thought to be an unknown moderator variable relative to one or more of the parameters in the regression model of interest. To apply the method, a researcher specifies the number of subpopulations, segments, or levels of the latent class variable that are thought to exist. Since this is rarely known, the regression mixture analysis typically is applied to multiple models where the models vary only in the number of classes of the latent class variable. An analyst might test a one class model, a two class model, a three class model, and so on and then make a judgment about which of the models best account for the data. The chosen model then dictates the number of levels of the unknown moderator, i.e., the unknown latent class variable. The mathematics of the approach permit the analyst to estimate the regression equation for each class in the chosen model, the proportion of people in each class, and the likelihood that a given person in the data is a member of each class. Consideration of the statistical theory underlying regression mixture modeling is complex and beyond the scope of this primer. Interested readers are referred to Magidson and Vermunt (2004).

Individuals in the different latent classes might differ from one another qualitatively and, accordingly, represent “true” subpopulations in the larger population, such as different ethnic groups or different religious affiliations. Alternatively, the different classes might approximate an underlying quantitative variable represented as discrete categories. As such, the classes might indicate either qualitative differences between individuals, quantitative differences between individuals, or a combination of the two (Bauer & Curran, 2003a, 2003b, 2004). Interpretation of the substantive meaning of the different classes should consider such possibilities and, ideally, evidence should be brought to bear to empirically validate the interpretations, as discussed below.

Regression mixture modeling makes the following assumptions: (1) the effect of the predictors on the outcome are linear (although this assumption can be relaxed), (2)

observations in the population are independent, (3) the population error scores are normally distributed within each class, and (4) the regression predictors have non-consequential measurement error. The magnitude of the error variances (the variance of the ε) can differ across the classes. Research suggests that violation of the normality assumption can bias parameter estimates, in which case remediation methods may be needed (Van Horn et al., 2012; George et al., 2013). For a comparison of regression mixture modeling to traditional product term analysis, see Van Horn et al. (2015).

INFORMATION INDICES AS A STRATEGY FOR MODEL CHOICE

When choosing between the different models to determine the number of classes, a commonly used set of comparative fit indices is based in a statistical theory known as *information theory*. Two such indices are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In general, researchers calculate an AIC index and/or a BIC index for the different models and then choose the model that has the best BIC or AIC value. In this section, we digress a bit and develop the logic of these indices, taking a few liberties in the interest of pedagogy. We first develop the concept of a log likelihood, a concept that is central to both the AIC and BIC. We then describe the model comparison process for the AIC, followed by a consideration of that process for the BIC.

Log Likelihoods

Suppose we have a very large population and half the population is male and half the population is female. The probability of a randomly selected case being a male is 0.50 and this also is true for being a female. Stated more formally:

p(male) = 0.50 p(female) = 0.50

If we randomly select two cases, the probability of a given joint result across the two selections or "trials" is the product of their probabilities. As such, the probability of observing two males is

p(male)*p(male) = (0.50)(0.50) = 0.25

This is known as the multiplication rule for independent trials. Stated more formally, let p(A) = the probability of event A on a trial and p(B) = the probability of event B on a second (independent) trial. The joint probability of both events A and B is the product of the individual probabilities p(A) p(B). To be more concrete, there are four combinations that can result, each with a probability of 0.25:

Probability of a male on the first trial followed by a male on the second trial: 0.25
Probability of a male on the first trial followed by a female on the second trial: 0.25
Probability of a female on the first trial followed by a male on the second trial: 0.25
Probability of a female on the first trial followed by a female on the second trial: 0.25

and if we do not care about the order of appearance in the trials,

Probability of two males: 0.25
Probability of a male and a female: 0.50
Probability of two females: 0.25

We now review another facet of statistical theory that we will make use of. If we know that a very large set of scores is normally distributed with a certain mean and standard deviation, then we can use knowledge of the probability density function for a normal distribution to compute the probability of obtaining any given value when we randomly select a case from that distribution. The density formula is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{\frac{-.5(x-\mu)^2}{\sigma^2}}$$

where x is the score value in question, μ is the mean of the distribution, σ is the standard deviation of the distribution, π is the mathematical constant pi, *e* is the constant associated with the Naperian logarithm, and the density describes the height of the normal curve at the value of x. We can use this density in conjunction with calculus to calculate the probability of observing the score in question. As an example, if scores are normally distributed with a mean of 100 and a standard deviation of 13.77, then, using the above formula, we find that the likelihood of a score of 99 is 0.0289. For a score of 87, it is 0.0186.[1]

Suppose we randomly select two scores from an extremely large population where scores are normally distributed with a mean of 100 and a standard deviation of 13.77. The probability that the scores will be 87 and 99, using the joint probability theorem described above, is (0.0289)(0.0186) = 0.00053754. Stated another way, the probability of observing these two data points given that the mean is 100 and the standard deviation

[1] Technically, the probability of observing an exact value for a continuous variable is zero. We compute the likelihoods here by focusing on the interval defined by the real limits of the number (e.g., 98.5 to 99.5) in conjunction with the integral that scales the area under the curve to 1.00.

is 13.77 (and assuming a normal distribution) is 0.00053754, with further adjustments to account for disinterest in the order of selection.

Suppose we randomly sample 100 data points from the population and calculate the likelihood of those 100 data points occurring using a strategy similar to the above method. The strategy would involve multiplying each probability by one another, with the result being a very, very small number. To make things more manageable and so as not to work with such small numbers, statisticians transform the final result by calculating the log of it, yielding what is called a *log likelihood*. The log likelihood is indicative of (but not equal to) the probability of obtaining the sample data given a "model" that states (a) the scores are normally distributed, (b) the mean is 100, and (c) the standard deviation is 13.77.

Log likelihoods are negative because the log of numbers less than 1.00 is always negative. For example, the natural log of 1.00 is zero, the natural log of 0.50 is -0.69, the natural log of 0.25 is -1.39, and the natural log of .01 is -4.61.[2]

Now, let's turn the above situation on its head. Suppose we have a set of 100 data points but we do not know the mean and standard deviation of the (assumed normal) distribution from which they come. We might, based on theory or logic, decide to "test" a model that states the mean is 95 and the standard deviation is 15. Using the probability density function from above and the strategies described, we can calculate the log likelihood for this model. The closer the log likelihood value is to zero (i.e., the less negative it is), the more likely the data came from the postulated model. We might formulate a second (competing) model that the mean is 100 and the standard deviation is 13.75 and calculate the log likelihood for it. Again, the closer the value of the log likelihood for this model is to zero, the more likely it is the data came from the model positing a mean of 100 and a standard deviation of 13.75.

We can compare the log likelihood values for the two models and we might find that one model results in a log likelihood closer to 0 than the other model. The model with the log likelihood closer to zero is more likely to have produced the data, hence we would prefer it to the model with the more negative log likelihood. Such is the fundamental logic of choosing between models based on their relative log likelihoods: We calculate the log likelihood of competing models and then choose the model with the log likelihood that is closest to zero. To be sure, the above explanation is simplistic and glosses over technicalities, but hopefully it conveys the general idea of comparing log likelihoods for two models.

As an aside, the above logic also is central to the well-known method of estimation

[2] Actually, some operationalizations of log likelihoods can yield positive numbers, but discussion of this point is beyond the scope of this primer.

called *maximum likelihood estimation*. In this approach, to estimate the mean of a distribution, one conceptually posits different models each representing a possible population mean value, calculates the likelihood of observing the data given the "model," and then selects the value/model that has the maximum likelihood.

Model Comparisons using the AIC

The AIC is an index of model likelihood or "model fit" based on a log likelihood. A common representation of it is

$$AIC = (-2)\,(LL) + 2k \qquad [1]$$

where LL is the log likelihood associated with the model in question and k is the number of estimable parameters in the model (such as when we estimate an intercept and the various regression coefficients). By multiplying the log likelihood by -2, the AIC essentially becomes a positive number, with larger numbers indicating lower likelihoods of the model. The AIC also includes what is often referred to as a penalty function for lack of parsimony, namely 2k. If the model has many parameters in it that must be estimated, then the AIC will be larger, everything else being equal. With the AIC, model parsimony is rewarded.[3] In general, the smaller the value of AIC, the better the "fit" of the model to the data. To make this intuitive, if the probability of the data given the model is 0.25, the log likelihood will be -1.39 and multiplying this by -2 yields 2.78. If the probability of the data given the model is much higher, say 0.50, the log likelihood is -0.69 and multiplying this by -2 yields 1.38. So, the smaller the value, the better the model. To this term, a penalty function is added that inflates the value of AIC for models that estimate more parameters.

There are many variations of the AIC. For example, some researchers use the above formula but with a small sample bias correction incorporated into it. This is sometimes referred to as AIC_c. The nuances of the different versions of the AIC are described in Burnham and Anderson (2004). Do not be surprised if for some software you observe AIC indices that are quite different in magnitude from other software. The important idea for all them is that we can compare different models using their respective AICs and then choose models that have "better" AICs when compared to other models.

Sometimes we compare more than two models, i.e., we might compare three, four or five models. When comparing more than two models, it is common to first identify the model with the lowest AIC value (which is the best fitting model of all the models being

[3] Technically, the 2k term is part of the mathematical theory underlying the derivation of AIC. Also, choosing the value of -2 to multiply the LL by is not arbitrary. This value has a clear rationale. See Burnham and Anderson (2004).

considered). One then calculates the difference in AIC values between each of the models and this best fitting model (subtracting the latter from the former). For the best fitting model, the difference will be zero and for all other models, it will be positive in value, with the larger the disparity, the worse the fit of the target model relative to the best fitting model.

General rules of thumb have been proposed to contextualize the magnitude of the difference in AICs between models (see Burnham & Anderson, 2004). The most common rules of thumb are as follows:

1. If the disparity in AICs is < 2, then the two models have about the same support

2. If the disparity in AICs is > 2 and < 4, then the better fitting model has positive support relative to the model it is compared with

3. If the disparity in AICs is > 4 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the disparity in AICs is > 10, then the better fitting model has very strong support relative to the model it is compared with.

Of course, one must be careful when applying rules of thumb like this because they may not apply in all contexts. Indeed, some analysts object to their specification, arguing that they can result in the same rigid and counterproductive use of a criterion like "$p < 0.05$" that plagues null hypothesis testing frameworks.

Another standard for comparing two models vis-a-vis the AIC is to examine what is called the *evidence ratio*. Let D = the AIC for the worse fitting model of the two models minus the AIC for the better fitting model of the two models (and let *e* be the traditional Naperian constant). The evidence ratio is defined as

$$ER = 1 / e^{(-D / 2)}$$

where ER stands for "evidence ratio." It indicates how much more likely the better fitting model is (given the data) than the worse fitting model (given the data). For example, if the AIC for the better fitting model is 100 and for the worse fitting model it is 102, then the evidence ratio is

$$1 / e^{-(102-100) / 2)} = 2.63$$

The better fitting model is 2.63 times more likely to have yielded the data than the model it is being compared with.

Finally, some researchers normalize AIC differences relative to all models being compared so that they sum to 1. These are called *Akaike weights* and indicate the "weight of evidence" in favor of a model relative to *all* models in the comparison set. Akaike weights are distinct from evidence ratios because Akaike weights are impacted by the particular set of models being compared when the number of models is greater than two. Let us first describe how Akaike weights are calculated and then we will make them more concrete with an example.

To calculate the Akaike weight, each model is assigned an index of its likelihood relative to that of the best fitting model using the value from the denominator of the evidence ratio, $e^{(-D/2)}$, as the index. Let T = the sum of the $e^{(-D/2)}$ values across all the models being considered. Then the Akaike weight for a given model is defined as

$e^{(-D/2)} / T$

The weight ranges from 0 to 1.00, with higher values favoring the model in question.

To make this concrete, suppose we fit five different models to a set of data. Here is a table with the AICs, the differences between the model AIC versus the model with the lowest AIC, and the Akaike weights (w):

Model	AIC	D	$e^{(-D/2)}$	$w = e^{(-D/2)}/T$
1	204	2	0.3678	0.2242
2	202	0	1.0000	0.6094
3	206	4	0.1353	0.0824
4	206	4	0.1353	0.0824
5	214	12	0.0024	0.0015
Sum			T = 1.6408	1.0000

The sum of the weights across all five models is 1.00. The weights represent a continuous measure of relative strength of evidence for each model. Each weight can be crudely interpreted as the probability that the model is the best model among the set. In the present case, the data support Model 2.

The basic idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in AICs, the evidence ratios, the Akaike weights, and the substantive meaning/logical coherence of the models, in order to choose the best one.

Model Comparisons using the BIC

We describe the logic of the BIC using the Schwartz BIC, which is formally defined as

$$BIC = -2\ LL + \ln(N)\ k \qquad [2]$$

where k = the number of estimable parameters in the model, N = the sample size, and LL = the model log likelihood. Like the AIC, the smaller the BIC, the better the model fit, everything else being equal. Like the AIC, there is a penalty function for lack of parsimony, but the penalty is different than the AIC. The penalty is somewhat harsher for the BIC as opposed to the AIC. There are other instantiations of the BIC, and we discuss these below. For current purposes, we use the Schwartz formulation.

Like the AIC, it is not uncommon for the model with the smallest BIC to be used as a reference point for comparing models, with a common practice being to calculate the difference between each model in the model set and the model with the best BIC, like we did for the AIC. For the best fitting model, this difference will be zero.

To evaluate models in terms of BIC differences, general rules of thumb are (see Raftery, 1995):

1. If the BIC disparity < 2.2, then the better fitting model and the model it is compared with have about the same support

2. If the BIC disparity > 2.2 and < 6, then the better fitting model has positive support relative to the model it is compared with

3. If the BIC disparity > 6 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the BIC disparity > 10 then the better fitting model has very strong support relative to the model it is compared with

For similar but slightly different standards, see Wasserman (1997).

One also can calculate what is called a *Bayes Factor* (BF) for each model relative to the best fitting model. It is defined as

$$BF = e^{(D'/2)}$$

where D' is the BIC difference between the target model and the best fitting model. The Bayes factor is the probability that the model with the lower BIC produced the data

divided by the probability the model in question produced the data. For example, a BF = 10 means it is 10 times more likely the model with the minimum BIC produced the data than the model in question.

Finally, a relative model weight, analogous to the Akaike weight, can be computed by normalizing model likelihoods relative to *all* models in the comparison set so that they sum to 1. Let D = the difference in the BIC for the model in question minus the value of the BIC for the best fitting model, T = the sum of the index $e^{(-D/2)}$ across each model. The relative weight for a model is

$$e^{(-D/2)} / T$$

The weight ranges from 0 to 1.00, with higher values favoring the model. Again, the sum of the weights across models is 1.00.

As with the AIC, the basic idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in BICs, the Bayes factors, the relative weights, and the substantive meaning/logical coherence of the models, in order to choose the best one.

You will encounter variants of the BIC, but the basic logic in applying them is the same. For example, like the AIC_c, there is a sample size adjusted BIC that is similar to Schwartz' BIC, but it applies a somewhat milder penalty function (Sclove, 1987). There also are variants of both the AIC and BIC to deal with dispersion issues in count regression models (called QAIC and QBIC).

Which Method is Better, AIC or BIC?

A debated topic in statistics is which approach to model comparison is better, one based on AICs or one based on BICs. There are advocates on both sides of the matter and we dare not venture into this controversy here. The BIC tends to favor simpler models more so than the AIC. This can be both a strength and a weakness. Interested readers are referred to Burnham and Anderson (2004), Yang (2005), and Kuha (2004). Kuha argues for the use of both indices.

An issue with both approaches is that researchers can be lulled into thinking that the best fitting model within a set of models is the true model. This is not necessarily the case. Researchers can choose the best of a set of wrong models, which is not our goal.

In regression mixture modeling, the choice of the number of latent classes for a model is often guided by the AIC and BIC values of the models with differing numbers of latent classes.

ADDITIONAL CRITERIA FOR MODEL CHOICE IN MIXTURE REGRESSION

In addition to the AIC and BIC, another consideration for evaluating model adequacy derives from what is known as a *classification analysis* for the mixture model. For each person in an analysis of a given model, regression mixture modeling provides an estimate of the probability that the individual is in each of the latent classes specified by the model. For example, in a three subgroup/class model, a given respondent will be assigned a probability that s/he is in subgroup 1, a probability that s/he is in subgroup 2, and a probability that s/he is in subgroup 3. The three probabilities for a given individual will sum to 1.0. The individual is classified into the subgroup that s/he has the highest probability of being in. A desirable model is one where individuals tend to have a high probability of being in one subgroup, but low probabilities of being in the other subgroups, i.e., the classification is well-differentiated. Here is an example classification analysis for a two subgroup mixture regression model:

RESPONDENTS PREDICTED TO BE IN SUBGROUP 1

Mean probability of being in group 1 = .92171
Mean probability of being in group 2 = .06653

RESPONDENTS PREDICTED TO BE IN SUBGROUP 2

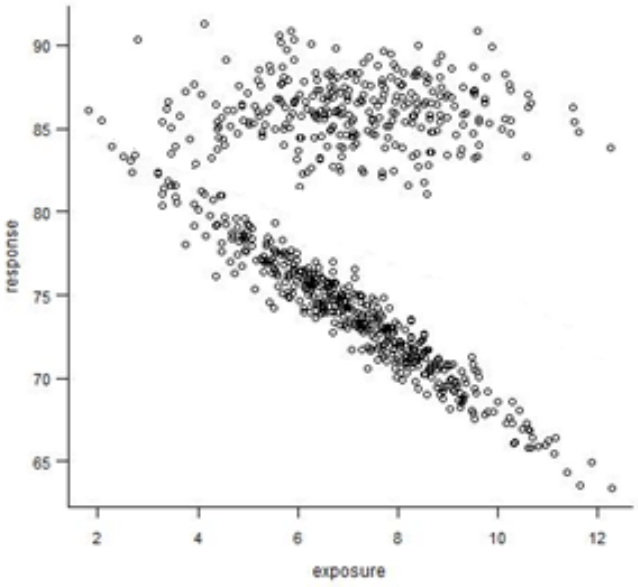
Mean probability of being in group 1 = .07829
Mean probability of being in group 2 = .93347

For those individuals assigned to subgroup 1, the average probability they had of being in subgroup 1 was 0.9217 whereas the average probability they were in subgroup 2 was only 0.06653. This is a well-differentiated pattern. The same is true for individuals assigned to subgroup 2. This classification tool is sometimes called a *confusion matrix*.

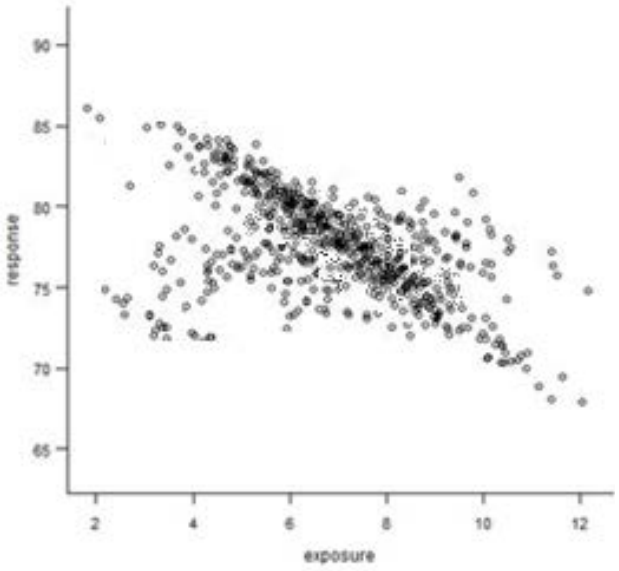
It is helpful to examine scatterplots where mixture modeling would yield well-differentiated versus less well-differentiated patterns in the confusion matrix. Figure 4.1 presents examples using a bivariate regression relating the degree of exposure to an environmental chemical (X) to beneficial response to a vaccine (Y). There are two segments in the data. Figure 4.1.a is well-differentiated because there is virtually no overlap between the two segments. Figure 4.1.b is less well-differentiated because of the overlap of the two segments in the middle of the plot. Overlapping individuals could be classified into either segment. Figure 4.1.c is well-differentiated because of non-overlap of the two segments but note that the regression coefficients for the two segments are

essentially the same (because the slopes are similar); it is the intercepts that differ (because one segment is elevated above the other segment).

(a) Well-differentiated pattern 1



(b) Less well-differentiated pattern



(c) Well-differentiated pattern 2

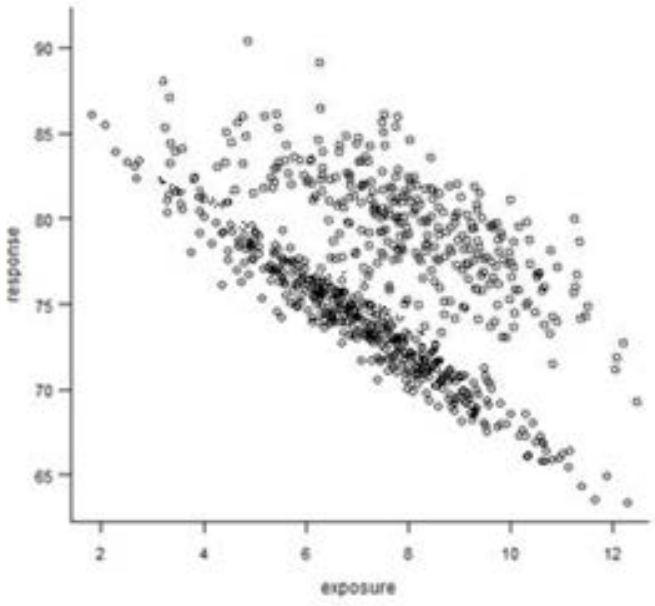


FIGURE 4.1. Examples of Differentiation of Segments

The dynamics of Figure 4.1.b are a reason some statisticians argue not to rely too heavily on confusion matrices to make decisions about the number of segments. There clearly are two distinct segments in this Figure and to combine them could be misleading. However, the confusion matrix will show a pattern that is not well-differentiated because there is a subgroup of individuals (in the middle) for whom classification is ambiguous. Such are the realities of the real world and, the argument goes, this is not a reason to reject the existence of the two segments.

Regression mixture models also provide estimates of the proportion of the population that is in each subgroup. This is useful because the size of the subgroup may be so small that the segment is of little interest. One might orient towards a group that represents 50% of the population differently than one that represents only 1% of the population. Sometimes an analyst is reluctant to move to a model with more classes if the new model produces additional classes that are very small in size.

Yet another criterion used by the analyst to settle upon a model is the substantive meaningfulness of adding more classes. If as a result of increasing the number of classes by 1, a particular subgroup is split into two subgroups that make more conceptual sense in a broader theoretical context, then one prefers the more complex model. However, if

the division that occurs by adding a class yields a class that makes no substantive sense, one might be reluctant to adopt the more complex model.

GIVING MEANING TO THE LATENT CLASSES

Once the number of classes is determined, regression mixture analysis reports the estimated regression equation for each class. These equations can help give substantive meaning to the classes. For example, a theory of vaccination behavior might hold that the intention (I) to vaccinate one's child against the measles is impacted by two classes of variables, (1) what a parent sees as the positives and negatives of having his or her child vaccinated (called the person's personal attitude (PA) toward getting a vaccination), and (2) the normative pressures from important others (N) the parent feels to obtain or not obtain a vaccination. Suppose all 3 constructs (I, PA and N) are measured on a 0 to 10 metric with higher scores favoring obtaining a vaccination. A regression mixture analysis might yield 3 classes/segments with the following regression equations (note: * indicates $p < 0.05$):

Segment 1: $I = 0.02 + 1.01^{*} PA + 0.01 N$
Segment 2: $I = 0.01 + 0.01 PA + 1.02^{*} N$
Segment 3: $I = 0.02 + 0.50^{*} PA + 0.50^{*} N$

The first segment is people whose intent to obtain a vaccination is primarily impacted by their personal attitudes. The second segment is people whose intent to obtain a vaccination is primarily impacted by the normative pressures of important others. The third segment is people whose intent to obtain a vaccination is impacted equally by both personal attitudes and normative pressures. This mixture analysis suggests different strategies for encouraging people to obtain vaccinations may be needed for different segments of the population; one strategy is needed to address personal attitudes for segment 1, a second strategy is needed to address normative pressures for segment 2, and both of these determinants need to be addressed for segment 3.

Suppose we learn from the regression mixture analysis that 62% of the population is in segment 1, 8% of the population is in segment 2, and 30% of the population is in segment 3. This information also might be useful for decisions about the logistics surrounding educational strategies about the vaccine. For example, the sizes of segments 1 and 3 might lead us to prioritize these groups. Also of interest is the mean intention for each of the classes. If the intent to obtain a vaccination is generally high for segment 2 but low for segments 1 and 3, that might be another reason for prioritizing segments 1 and 3.

Another type of analysis that can give meaning to the classes is to identify correlates of class membership external to the mixture analysis. For example, we might find that membership in the second segment is correlated with elevated scores on an impression management scale, being female, being older, and having anxiety, all of which past research has shown relate to conformity. Such analyses should take into account the probabilistic nature of group membership, perhaps through the use of weights based on those probabilities (Bakk, Oberski & Vermunt, 2014).

THE EXPLORATORY NATURE OF REGRESSION MIXTURE MODELING

The main text of Chapter 11 discussed regression mixture modeling as an exploratory method of data analysis to facilitate theory construction. Because of its exploratory nature, in more traditional research contexts, it generally is useful to replicate one's results from regression mixture modeling with independent samples to increase one's confidence in segment identification.

CONCLUDING COMMENTS

When making conclusions from multiple regression analysis, we often implicitly assume the individuals studied represent a single population with common parameter values in the equation. It is possible, however, that the population is composed of mixtures of sub-populations that have different regression coefficients and/or a different intercept, that is the population is a mixture of two or more heterogeneous segments with distinct coefficients. In such cases, the population regression model can be misleading and mischaracterize the separate population segments. Regression mixture modeling is an exploratory method for detecting the presence of heterogeneous segments in a population. It conceptualizes the different segments as an unknown categorical latent class variable that moderates the effect of predictors on the outcome. Application of regression mixture modeling requires comparative tests of models that vary in the hypothesized number of distinct segments in the population. These comparative tests often make use of the Akaike Information Criterion and the Bayesian Information Criterion. As well, researchers examine a confusion matrix, the size of the segments, and the substantive meaning of the segments to make decisions about the most plausible number of segments to focus on. Regression mixture modeling is a useful exploratory method for potentially enriching theory surrounding moderated relationships.

REFERENCES

Bakk, Z., Oberski, D. & Vermunt, J. (2014). Relating latent class assignments to external variables: Standard errors for correct inference. *Political Analysis*, 22, 520-540.

Bauer, D. J., & Curran, P. J. (2003a). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, 8, 338-363.

Bauer, D. J., & Curran, P. J. (2003b). Overextraction of latent trajectory classes: Much Ado about nothing? Reply to Rindskopf (2003), Muthen (2003), and Cudeck and Henly (2003). *Psychological Methods*, 8, 384-393.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, 9, 3-29.

Burnham, K. & Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33, 261-304.

George, M., Yang, N., Jaki, T., Feaster, D., Lamont, A., Wilson, D. & Van Horn, M.L. (2013). Finite mixtures for simultaneously modelling differential effects and non-normal distributions. *Multivariate Behavioral Research*, 48, 816–844.

Kuha, J. (2004). AIC and BIC : Comparisons of assumptions and performance. *Sociological Methods & Research*, 33, 188-208.

Lazarsfeld, P.F. & Henry, N.W. (1968) *Latent structure analysis*. Boston: Houghton Mifflin

Magidson, J., & Vermunt, J. K. (2004). *Latent class models*. Thousand Oaks: Sage.

McCutcheon, A.L. (1987). *Latent class analysis*. Thousand Oaks, CA: Sage

Muthén, B. O., & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society, Series A*, 172, 639-657.

Raftery, A.E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*, 25, 111-195.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika,* 2, 333-343

Van Horn, M. L., Smith, J., Fagan, A. A., Jaki, T., Feaster, D., Masyn, K., et al. (2012). Not quite normal: Consequences of violating the assumption of normality with regression mixture models. *Structural Equation Modeling,* 19, 227-249.

Van Horn, M.L., Jaki, T., Masyn, K., Howe, G., Feaster, D., Lamont, A., George, M. & Kim, M.(2015). Evaluating differential effects using regression interactions and regression mixture models. *Educational and Psychological Measurement,* 75, 677-714.

Wasserman, L. (1997). Bayesian model selection and model averaging (Working Paper No. 666). Carnegie Mellon University, Department of Statistics.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika,* 92, 937-950.