4

Worked Example for Regression Mixture Analysis

This example uses the ASA software integrated into Excel or SPSS (www.asastat.com). ASA is, in part, a point-and-click interface to R but analyses can be conducted from within SPSS or Excel. All data are hypothetical. We assume you have read the primer on mixture regression.

The example focuses on the relationship between naturally occurring exposure to a chemical that is thought to inhibit response to a vaccine. The index of response to the vaccine is measured as a percentage and ranges from 0 to 100 with higher scores indicating a better response. Exposure levels to the chemical range from 1.8mg to 12.5mg and are in the variable called *exposure*. Vaccine response is the variable called *response*.

The ASA software routinely reports confidence intervals for key parameters in statistical models. There are different ways of presenting confidence intervals. One strategy is to report them directly. Another strategy is to report them as margins of error, much like the margins of error you see for political polls on television or in print media. In this case, one calculates the half width of the confidence interval and reports it in "plus or minus" format. For example, in a political poll, you might be told that the percent of people endorsing a candidate is 50% ±5%. In this case, the confidence interval is 45% to 55%. This is an efficient way of summarizing the interval. In some cases, confidence intervals are asymmetric. When this occurs, some researchers will report the lower and upper margin of error separately. Alternatively, the researcher might calculate the absolute difference between the lower limit and the parameter estimate as well as the absolute difference between upper limit of the interval minus the parameter estimate and then report whichever difference is larger using the ± format. Some analysts prefer the use of credible intervals in Bayesian analytic frameworks instead of confidence intervals for characterizing margins of error (see Curran, 2005).

The mixture regression program is called "Mixture regression" and is located in the ASA folder "Multiple Regression: Interaction Analysis > Mixture Regression" (see the video on our website). Mixture regression is used to evaluate a series of models to determine how many heterogeneous subgroups or segments are present in the population relative to the bivariate regression model that regresses vaccine response onto exposure. We specify in the ASA program that we want to evaluate a two class model, so the program, by default, tests the fit of a one class model, a two class model, a three class

model, and a four class model, i.e., it evaluates a range of models around the *a priori* specified number of classes. The comparisons use the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). If you are unfamiliar with these indices, see the primer on regression mixture models. Here is the relevant output:

```
MODEL EVALUATION WITH ALTERNATIVE NUMBER OF SUBGROUPS

NUMBER OF SUBGROUPS: 1

   Log likelihood: -2461.91500
   AIC:            4929.83100
   BIC:            4943.69000

NUMBER OF SUBGROUPS: 2

   Log likelihood: -1742.15100
   AIC:            3498.30200
   BIC:            3530.64200

NUMBER OF SUBGROUPS: 3

   Log likelihood: -1741.92800
   AIC:            3505.85500
   BIC:            3556.67600

NUMBER OF SUBGROUPS: 4

   Log likelihood: -1736.93200
   AIC:            3503.86500
   BIC:            3573.16600
```

The model with the lowest AIC and the lowest BIC is the best fitting model, which is the model with two subgroups. To formalize the model comparisons, we use the ASA program called “Model comparison using AIC and BIC” in the folder “Model Fit” to more formally compare the four models. We entered the values of the AIC and the BIC provided by the output for input into this program. We initially focus on the results for the AIC. The model comparison program first identifies the model with the minimum AIC value:

```
Model with minimum AIC: Model 2
Minimum AIC value: 3498.302
Akaike weight for model with minimum AIC: .9218
```

The Akaike weight ranges from 0 to 1.00 and is an index of the degree to which the results favor the minimum AIC model over the other models. The closer the value is to 1.00, the more the minimum AIC model is favored. It can be crudely interpreted as the probability that the model is the best model among the set of models. A value of 0.92 for

the two class model is quite large. Here is the additional comparative information from the output:

```
ANALYSIS OF AICs

Model 1 compared with minimum AIC model (Model 2)

   Model AIC: 4929.831
   AIC difference: 1431.529
   Evidence ratio: > 1,000
   Akaike weight: .0000

Model 3 compared with minimum AIC model (Model 2)

   Model AIC: 3505.855
   AIC difference: 7.553
   Evidence ratio: 43.663
   Akaike weight: .0211

Model 4 compared with minimum AIC model (Model 2)

   Model AIC: 3503.865
   AIC difference: 5.563
   Evidence ratio: 16.143
   Akaike weight: .0571
```

The Akaike weights are much smaller for the competing models, thereby favoring the two class model. The AIC differences are simply the difference between a given model and the best fitting model; the larger the difference, the worse the target model accounts for the data relative to the best fitting model (Model 2). Model 1 (the one class model) is clearly not in the mix given the sizeable difference and using the comparison criteria of Burnham and Anderson (2004). The two class model yielded a 5.563 lower AIC value than the four class model. Based on Burnham and Anderson (2004), this indicates strong support for the two class model over the four class model. The support for the two class model over the three class model is even stronger as reflected by a difference of 7.553. The evidence ratio indicates how much more likely the best fitting model (Model 2) is than the target model for having been the source of the data. For example, for Model 4, the evidence ratio is 16.14, which means Model 2 is 16.14 times more likely to have generated the data than Model 4. All of the indices, taken together, clearly point to Model 2 (two segments) as the best model.

The results for the BIC were comparable to those for the AIC. Here is the output:

```
Model with minimum BIC: Model 2
Minimum BIC value: 3530.642
Relative weight for model with minimum BIC: 1.0000
```

```
Model 1 compared with minimum BIC model (Model 2)

   Model BIC: 4943.690
   BIC difference: 1413.048
   Bayes factor: > 1,000
   Relative model weight: .0000

Model 3 compared with minimum BIC model (Model 2)

   Model BIC: 3556.676
   BIC difference: 26.034
   Bayes factor: > 1,000
   Relative model weight: .0000

Model 4 compared with minimum BIC model (Model 2)

   Model BIC: 3573.166
   BIC difference: 42.524
   Bayes factor: > 1,000
   Relative model weight: .0000
```

Model 2 had the lowest BIC. The relative weights for the competing models were near zero (although they show a value of zero because ASA only reports results to 4 decimals) and the relative weight for Model 2 was near 1.00. These weights are interpreted much like Akaike weights and favor Model 2. The BIC difference entries are the difference between a given model and the best fitting model; the larger the difference, the worse the target model accounts for the data relative to the best fitting model (Model 2). Model 1 (the one class model) again is clearly not in the mix given the large value of the difference and based on the criteria of Raftery (1995). The two class model yielded a BIC that was 26.034 units lower than the BIC for the three class model. Based on Raftery (1995), this indicates very strong support for the two class model over the three class model. The support for the two class model over the four class model is even stronger. The Bayes Factor indicates how much more likely the best fitting model (Model 2) is than the target model in terms of being the source of the data. For example, for Model 3, the Bayes Factor is larger than 1,000, clearly indicating the superiority of Model 2 to the three class model. All of the indices, taken together, again point to Model 2 (two segments) as the best model, so we settle upon it for further exploration and analysis.

ASA also provides information about how clearly differentiated the subgroup differences are at the individual level using a confusion matrix. For each individual, the program estimates the probability the individual is in subgroup 1 and the probability the individual is in subgroup 2. The individual is classified as belonging to the subgroup that has the higher of these probabilities. For example, if the probability the person is in subgroup 1 is 0.92 and the corresponding probability that the person is in subgroup 2 is 0.08, then it makes sense to classify the individual as a member of subgroup 1. If the data

are well-differentiated, one would expect the average probability of being in group 1 to be larger for those classified into subgroup 1 and the average probability for being in subgroup 2 to be small. Here is result:

```
CLASSIFICATION ANALYSIS

   RESPONDENTS PREDICTED TO BE IN SUBGROUP 1

      Mean probability of being in group 1 = .99777
      Mean probability of being in group 2 = .00488
```

This indicates a well-differentiated classification structure. Here is the corresponding results for individuals classified into subgroup 2:

```
   RESPONDENTS PREDICTED TO BE IN SUBGROUP 2

      Mean probability of being in group 1 = .00223
      Mean probability of being in group 2 = .99512
```

Again, the structure of the data is well-differentiated.

Next, we examine the regression equations for the two subgroups, which is provided on the output. Here is the regression equation for the first subgroup identified by the program, beginning with the intercept:

```
REGRESSION EQUATIONS FOR SUBGROUPS

Subgroup 1

   Value of intercept: 88.1902
   Standard error: .1956
   95% confidence interval: 87.8068 to 88.5736
   Margin of Error: +/- .383
   z value: 450.8253
   p value: .000000
```

The intercept for this group is the predicted mean vaccine response when exposure is 0. This is 88.19 ± 0.38. However, because an exposure of 0 is outside the range of exposure levels in the data, we should be cautious about generalizing to this value. Given this, we do not consider the intercepts further.

Here is the regression coefficient for the first subgroup:

```
   Regression coefficient for EXPOSURE: -2.0202
   Standard error: .0265
   95% confidence interval: -2.0722 to -1.9682
   Margin of Error: +/- .052
   z value: -76.1005
   p value: .000000
```

The value of -2.02 indicates that for every unit that exposure increases, the mean response to the vaccine is predicted to decrease by 2.02 percent ±0.05. The coefficient is statistically significant ($z = 76.10$, $p < 0.05$).

Here is the intercept for the second subgroup:

```
Value of intercept: 86.1626
Standard error: .4796
95% confidence interval: 85.2226 to 87.1027
Margin of Error: +/- .940
z value: 179.6511
p value: .000000
```

Here is the regression coefficient for the second subgroup:

```
Regression coefficient for EXPOSURE: -.0068
Standard error: .0649
95% confidence interval: -.1341 to .1204
Margin of Error: +/- .127
z value: -.1052
p value: .916236
```

The value of -0.007 was not statistically significant ($z = 0.11$, *ns*).

It appears that vaccine response for one segment of the population is relatively unaffected by the degree of exposure to the chemical but the other segment is affected by it. We flush this out in more depth below.

The program also reports the error variances for the regression equations for the two classes, in standard deviation form:

```
ERROR VARIANCES FOR SUBGROUPS

Subgroup 1, error standard deviation: .9593
Subgroup 2, error standard deviation: 1.9262
```

These statistics give us a sense of how well the exposure levels predict the response to the vaccine in each subgroup. Some researchers prefer a squared multiple correlation but this index also is intuitive. The error-based standard deviations can be interpreted as the average disparity between the observed and predicted outcome scores for each group. For the first subgroup, the average error in prediction was 0.96 (on a metric of 0 to 100 percentage points given the response was measured in percent units) and for the second subgroup it was 1.93. Both seem quite respectable.

Finally, as part of the standard output, the program estimates the proportion of the population that is in each subgroup:

```
ESTIMATED PROPORTION OF POPULATION IN EACH SUBGROUP

Subgroup 1: .5960
Subgroup 2: .4040
```

About 59.6% of the population is estimated to be in the first subgroup, and 40.4% of the population is estimated to be in the second group. If the size of a subgroup is small, then it may be of less interest substantively.

In terms of journal write-up, space constraints dictated by journals typically restrict providing too much detail about the analysis. Here is how we might write-up these results for a report assuming we have already dealt with the issues of regression assumptions and have explained how we defining margins of errors (e.g., “Margins of errors (MOEs) are calculated from 95% confidence intervals and are the absolute distance between the lower limit or upper limit of the interval minus the parameter estimate, whichever is larger, unless otherwise noted”):

“A regression mixture model was fit to the data for exposure and response to the vaccine. Model comparisons were made for a one group, two subgroup, three subgroup, and four subgroup solution. The respective log likelihoods for these models were -2,461.92 (Akaike Information Criterion (AIC) = 4,929.83, Bayesian Information Criterion (BIC) = 4,943.69), -1,742.15 (AIC = 3,498.30, BIC = 3,530.64), -1,741.93 (AIC = 3,505.86, BIC = 3,556.68), and -1,736.93 (AIC = 3,503.86, BIC = 3,573.17). The data clearly supported the two subgroup solution. For example, the Akaike weight for the two subgroup model was 0.92 as compared to 0.06 for the next best fitting model and its evidence ratio was 16.14 relative to the next best fitting model. A classification analysis for the two group solution found that for individuals classified into subgroup 1, the average probability of being a member of group 1 was 0.998 and the average probability of being a member of group 2 was 0.005. For individuals classified into subgroup 2, the corresponding average probabilities were 0.002 and 0.995. These results indicate a well-defined data structure for the two groups.

For subgroup 1, the regression coefficient for exposure was -2.02 (±0.05, z = 76.10, $p < 0.05$), indicating that for every unit increase in exposure, the mean response to the vaccine was predicted to decrease by 2.02 percent. For the second subgroup, the regression coefficient was -0.007 (±0.13, z = 0.11, *ns*). The estimated percent of individuals in the first subgroup was 59.6% and for the second subgroup it was 40.4%. For the first subgroup, the average error in prediction as indicated by the square root of the mean square residual was 0.96 and for the second subgroup it was 1.93.”

As noted in the primer, we often are interested in identifying correlates of subgroup membership. To do so correctly, it is best to use state-of-the-art methods in the Mplus SEM software that take into account the probablisitic nature of group assignment. However, approximate exploratory analyses can be conducted that use the group membership information produced by the ASA software. ASA outputs for each individual the probability s/he is in each of the subgroups as well as the "assigned" subgroup of the individual, namely the subgroup the person has the largest probability of being in. This information can be saved in a data file and then integrated into your existing data. You can then perform exploratory analyses as a function of group membership. This strategy works best when there is a well-differentiated confusion matrix.

For the current example, we calculated the mean response to the vaccination for the two subgroups and found it to be 73.8 for the first segment and 86.2 for the second segment, a rather sizeable difference (Cohen's d = 4.18). The second segment, which was relatively unaffected by exposure, thus tended to show overall good response to the vaccine (86.2) but the segment that was affected by exposure showed an overall lower response to the vaccine (73.8). The mean degree of exposure to the chemical was about the same in the two segments (7.13 and 7.12, respectively). We also found that 66% of the individuals in segment 1 were Blacks but only 28% of the individuals in segment 2 were Black, suggesting ethnic differences in sensitivity to chemical exposure.

REFERENCES

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

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