Obtaining Meaningful Latent Class Segments with Ratings Data by Adjusting for Level and Scale Effects

by

Jay Magidson, Ph.D.

Statistical Innovations

**Abstract**

Individuals tend to differ in their use of rating scales in ways that confound the ability to obtain meaningful segments. With a 9-point rating scale for example, one person may rate liking for each of 15 food products in the 6-9 range for while another with similar preferences may provide ratings in the 4-7 range (level effect). Also, more conservative individuals might avoid giving the highest *and* lowest ratings while others might avoid use of mid-level ratings (scale effect). With traditional methods resulting segments may differ primarily in their rating response styles rather than their preferences.

This paper introduces an extended latent class (LC) model, available in the syntax module of Latent GOLD 5.0, that classifies individuals while also adjusting for level and scale effects. Applying this model to data from a taste-testing experiment, yields segments differing in more meaningful ways than those from more traditional LC models.

**Background**

It is well known that analysis of ratings data is difficult because different individuals have different response styles. For example with a 9-point rating scale, one person may give ratings in the 6-9 range with a mean rating of 7.5 for each of 15 food products while another with similar preferences may use only ratings in the 4-7 range with a mean rating of 6 (level effect). Also, more conservative individuals might avoid giving the highest and lowest ratings altogether while others might avoid use of mid-level ratings (scale effect). In such cases traditional clustering methods may yield segments that differ primarily in their rating response styles rather than their preferences.

In a food industry taste testing application Magidson, et al. (2006) showed that traditional latent class (LC) regression analysis resulted in 2 segments that differed only in their response styles – persons in class 1 tended to rate all crackers lower than those in class 2 (Figure 1) – a result that was not useful to the food manufacturer interested in developing different products for each segment based on their taste preferences. Use of a random intercept latent class (LC) regression model to control for level effects, resulted in segments that differed in their relative preferences, class 1 rating crackers #342 and #608 significantly higher than class 2 (Figure 2).

**Adjusting for scale effects**

For the purpose of this paper, we reanalyze the cracker ratings data using new models allowing for log-scale effects in addition to level effects (models 3 and 4 below). These log-scale models were proposed by Vermunt (2013) and implemented in the syntax module of Latent GOLD version 5.0. Specifically, scale classes (sClasses), which are latent classes that differ in their scale factor, are included in the model in addition to traditional classes, and for identification, one sClass is chosen as the reference and assigned a scale factor (SFactor) of 1. SFactors associated with each of the other sClasses are estimated simultaneously with the other model parameters.

Figure 1. Mean Liking Rating for each of 15 Cracker Products for 2 Latent Class Segments obtained from a Traditional LC Regression Analysis.

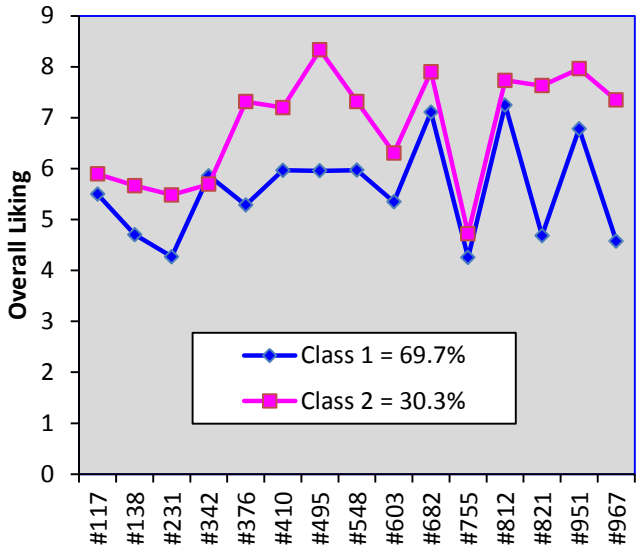



Figure 2. Mean Liking Rating for each of 15 Cracker Products for 2 Latent Class Segments obtained from a Random Intercept LC Regression Analysis.

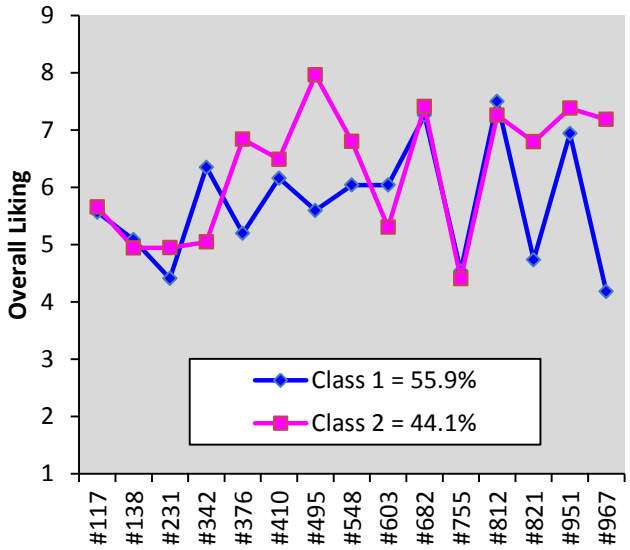
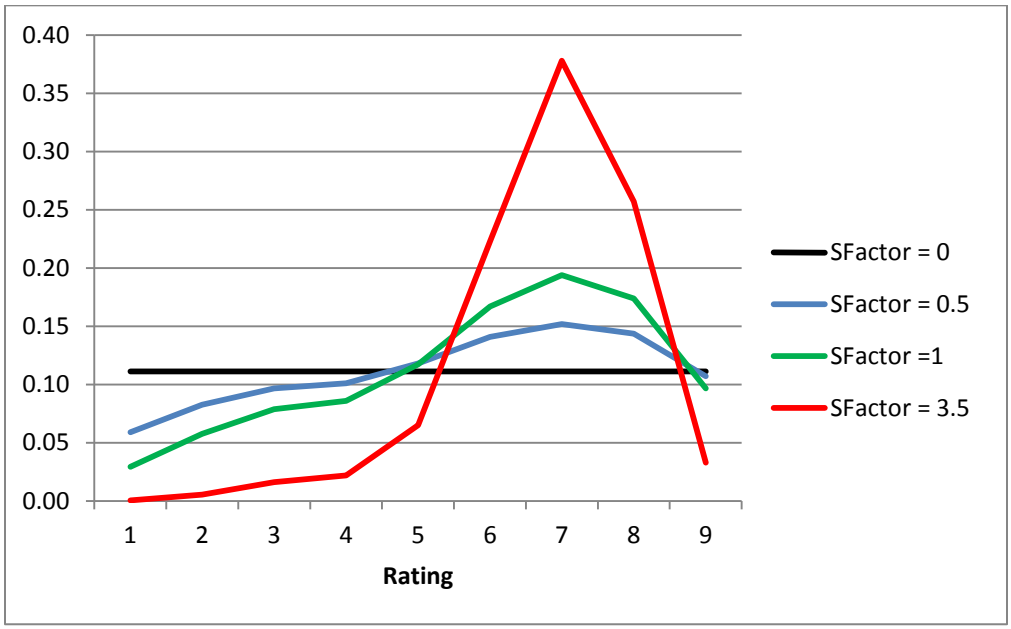



Figure 3 shows how a given ratings distribution becomes more extreme as the associated SFactor increases. For example, the green line plots the ratings level probabilities associated with the reference SClass (SFactor = 1), where the rating levels 6-8 have the highest probabilities of being used (prob > .16), with the log-odds of a '7' vs. a '6' rating is ln(.19/.17) = 0.15. The red line, plotting the probabilities

associated with SFactor = 3.5, has more extreme probabilities than those associated with SFactor = 1, the odds of a '7' vs. a '6' rating is now 3.5 times as high -- ln(.38/.22)=0.53.

Figure 3. Probabilities of Ratings for each of 4 Scale Factors (SFactors).




The scale factor is inversely related to the standard deviation of the ratings. Table 1 shows that the standard deviation is highest (2.6) under complete uncertainty (SFactor = 0). This situation is depicted by the black line in Figure 3 where each of the 9 rating levels are equally likely (prob = 1/9) of being chosen. The standard deviation steadily declines from 2.6 to 1.25 as SFactor increases.

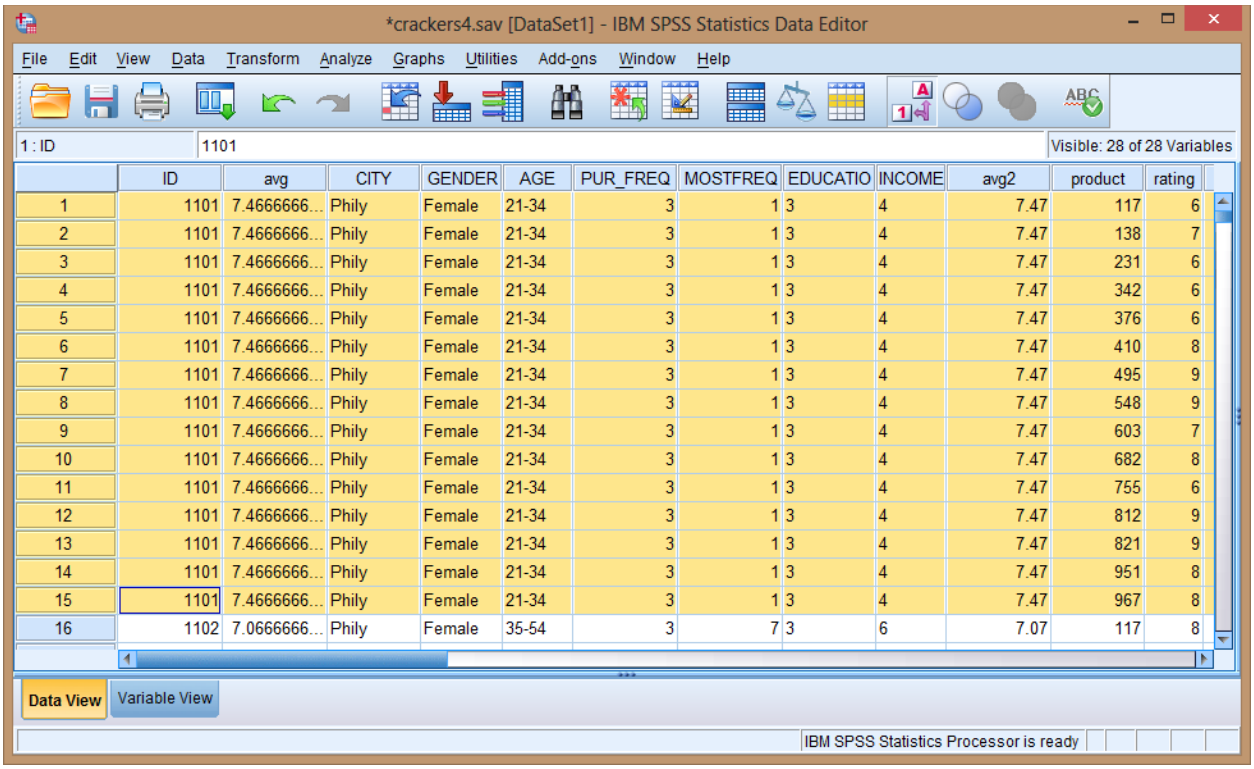
Table 1. Probability distribution and standard deviation for Rating associated with 4 hypothetical scale factors (SFactors)

| | Scale Factor | | | |
|---|---|---|---|---|
| **Rating** | **0** | **0.5** | **1** | **3.5** |
| 1 | 0.11 | 0.06 | 0.03 | 0.00 |
| 2 | 0.11 | 0.08 | 0.06 | 0.01 |
| 3 | 0.11 | 0.10 | 0.08 | 0.02 |
| 4 | 0.11 | 0.10 | 0.09 | 0.02 |
| 5 | 0.11 | 0.12 | 0.12 | 0.06 |
| 6 | 0.11 | 0.14 | 0.17 | 0.22 |
| 7 | 0.11 | 0.15 | 0.19 | 0.38 |
| 8 | 0.11 | 0.14 | 0.17 | 0.26 |
| 9 | 0.11 | 0.11 | 0.10 | 0.03 |
| **StdDev =** | 2.60 | 2.39 | 2.16 | 1.25 |

**Models**

For our LC segmentation analyses of the crackers data, we use the long file format (Figure 4) which allows us to specify the model as a LC regression model with the ordinal RATING variable as a function of the nominal variable PRODUCT as the sole observed predictor.

Figure 4. Crackers Ratings Data in Long File Format with 15 Records per Case.




| | ID | avg | CITY | GENDER | AGE | PUR_FREQ | MOSTFREQ | EDUCATIO | INCOME | avg2 | product | rating |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 117 | 6 |
| 2 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 138 | 7 |
| 3 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 231 | 6 |
| 4 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 342 | 6 |
| 5 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 376 | 6 |
| 6 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 410 | 8 |
| 7 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 495 | 9 |
| 8 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 548 | 9 |
| 9 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 603 | 7 |
| 10 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 682 | 8 |
| 11 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 755 | 6 |
| 12 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 812 | 9 |
| 13 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 821 | 9 |
| 14 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 951 | 8 |
| 15 | 1101 | 7.4666666... | Phily | Female | 21-34 | 3 | 1 | 3 | 4 | 7.47 | 967 | 8 |
| 16 | 1102 | 7.0666666... | Phily | Female | 35-54 | 3 | 7 | 3 | 6 | 7.07 | 117 | 8 |



We use the adjacent category logit model structure to formulate the various LC regression models with X denoting the nominal latent variable to define the segments explaining the heterogeneity in the PRODUCT effect. The models of interest have the following form:

Model A: Simple LC regression model with K classes

$$log\left[\frac{P(Y_{it}=m\,|\,x)}{P(Y_{it}=m-1\,|\,x)}\right]=\alpha_m+\beta_x+\gamma_{xt}$$ m=1,2,…,9 ratings, t=1,2,…,15 products, x=1,2,…K classes

In models containing scale classes, we estimate the *logarithm* of the scale factor to assure that the scale factor is non-negative. Taking $\lambda_j$ as the natural logarithm of the scale factor, for J scale classes we have:

$$log\left[\frac{P(Y_{it}=m\,|\,x)}{P(Y_{it}=m-1\,|\,x)}\right]=\exp(\lambda_j)(\alpha_m+\beta_x+\gamma_{xt})$$ j=1,2,…J sClasses (Model B)

where one of the log-scale factors is restricted to 0 (for identification).

To allow for the individual level effect, a random intercept F is added as an additional term. For example, for model A we have:

$$log\left[\frac{P(Y_{it}=m\mid x)}{P(Y_{it}=m-1\mid x)}\right]=\alpha_m+\mu F_i+\beta_x+\gamma_{xt}$$

while model B becomes:

$$log\left[\frac{P(Y_{it}=m\mid x)}{P(Y_{it}=m-1\mid x)}\right]=\exp(\lambda_j)(\alpha_m+\mu F_i+\beta_x+\gamma_{xt})$$

As an alternative to sCLasses, separate continuous factors (sCFactors) can be estimated for each individual (see Models 6 and 7 below).

Results are summarized below in Table 2. The best models are Model 5 with 2 scale classes and Model 7 with 1 sCFactor. Note that these models have almost identical log-likelihood values, Model 5 containing 1 additional parameter associated with the class sizes for the sClasses.

Table 2. Summary Results for Various Models Estimated

| Model | Description | LL | BIC(LL) | Npar |
|---|---|---|---|---|
| 1 | Ordinal Regression | -4762.4 | 9636.1 | 22 |
| 2 | 2-class Simple Ordinal Regression | -4682.4 | 9556.9 | 38 |
| 3 | 2-classes + Random Intercept | -4641.4 | 9480.1 | 39 |
| 4 | 2-cl w/Random Intercept + 2 scale classes | -4631.6 | 9470.6 | 41 |
| 5 | Model 4 + (class, sClass) correlation | -4626.8 | 9465.9 | 42 |
| 6 | 2-cl w/Random Intercept + scale CFactor | -4630.9 | 9464.0 | 40 |
| 7 | Model 6 + (class, sCFactor) correlation | -4626.4 | 9460.2 | 41 |

See the appendix for the Latent GOLD 5.0 syntax specification of these models.

## References

Magidson, J., J.K. Vermunt. (2006). "Use of latent class regression models with a random intercept to remove overall response level effects in ratings data". In A. Rizzi and M Vichi (eds.), *Proceedings in Computational Statistics* , 351-360, Heidelberg: Springer.

Vermunt, J.K. (2013). Categorical response data. In: M.A. Scott, J.S. Simonoff, and B.D. Marx (eds.), *The SAGE Handbook of Multilevel Modeling*, 287-298. Thousand Oaks, CA: Sage.

## Appendix

Rating is defined as an ordinal response variable using the adjacent category regression model. In all models, 2-classes are defined using the latent variable 'Class'

**Model 1: Ordinal Regression**

```
variables
  caseid ID;
  dependent rating;
  independent product nominal, avg inactive;
equations
  rating <- 1  + product ;
```

**Model 2: 2-class Simple Ordinal Regression**

```
variables
  caseid ID;
  dependent rating;
  independent product nominal, avg inactive;
  latent
    Class nominal 2;
equations
  Class <- 1;
  rating <- 1 + Class + product | Class;
```

**Model 3: 2-classes + Random Intercept**

```
variables
  caseid ID;
  dependent rating;
  independent product nominal, avg inactive;
  latent
    CFactor1  continuous,
    Class nominal 2;
equations
  (1) CFactor1 ;
  Class <- 1;
  rating <- 1 + Class + CFactor1 + product | Class;
```

**Model 4: 2-cl w/Random Intercept + 2 scale classes**

```
variables
   caseid ID;
   dependent rating;
   independent product nominal, avg inactive;
   latent
      CFactor1  continuous,
      Class nominal 2, sclass nominal 2 coding=first;
equations
   (1) CFactor1 ;
   Class <- 1;
  sClass <- 1;

   rating <- 1 + Class + CFactor1 + product | Class;
rating <<- sclass;
```

**Model 5: Model 4 + (class, sClass) correlation**

```
variables
   caseid ID;
   dependent rating;
   independent product nominal, avg inactive;
   latent
      CFactor1  continuous,
      Class nominal 2, sclass nominal 2 coding=first;
equations
   (1) CFactor1 ;
   Class <- 1;
  sClass <- 1;

   rating <- 1 + Class + CFactor1 + product | Class;
rating <<- sclass;
Class <->  sClass;
```

**Model 6: 2-cl w/Random Intercept + scale CFactor**

```
variables
   caseid ID;
   dependent rating;
   independent product nominal, avg inactive;
   latent
      CFactor1  continuous,
      Class nominal 2, scfac continuous;
equations
   (1) CFactor1 ;
 sCFac ;
   Class <- 1;

   rating <- 1 + Class + CFactor1 + product | Class;
rating <<- (1)scfac;
```

**Model 7: Model 6 + (class, sCFactor) correlation**

```
variables
   caseid ID;
   dependent rating;
   independent product nominal, avg inactive;
   latent
      CFactor1  continuous,
      Class nominal 2, scfac continuous;
equations
   (1) CFactor1 ;
 sCFac ;
   Class <- 1;

   rating <- 1 + Class + CFactor1 + product | Class;
rating <<- (1)scfac;
scfac <- class;
```