Supplementary Material

Kevin Jin

2024-01-17

Simulations

Simulation for Ordinal Response Variables with Multiple Levels

Let's now use simulations to check the efficacy this proposed method for visualizing the dependence structure and interactions of categorical data with an ordinal response which has multiple levels. We begin with visualizing interactions of categorical data for ordinal responses with multiple levels. We will begin with the cases of having three variables (response variable Y, explanatory variable X, and grouping variable G) and consider examples with more than three variables at the end of this section.

In our simulations with 3 variables, we will consider a hypothetical data set where the ordinal response Y has five categories $(y_1, \ldots, y_5)$, ordinal predictor X has three categories (x_1, x_2, x_3), and nominal predictor G has two categories (g_1, g_2) unless otherwise specified. While these number of categories were arbitrarily chosen, these combination of category number for each variable will be sufficient to consider multiple different relationships between the variables. Further, our BECCR predictions will always be done with a bootstrap resampling of $B = 1000$ to ensure consistency[1].

Simulating Hypothetical Data with Linear Relationship

For our first simulation, consider the following aggregated hypothetical data set shown in Figures 1 (mosaic plot) and 2 (BECCR Prediction Bubble Plot). We will consider this aggregate visualization first, as we want to see what kind of "Simpson's Paradox" may occur before accounting for grouping variable G. This will allow us to gain a better insight on how the BECCR Prediction Bubble Plots can be used to potentially visualize these amalgamation effects.

Looking at these visualizations, both the BECCR Prediction Bubble Plot and the mosaic plot do not seem to indicate that there is an relationship between X and Y. The hypothetical data set used is displayed in Table 1. From this, we can see, that after accounting for G, there appears to be a linear relationship between X and Y, with it being negative or positive depending on the level of G. In Figures 3 and 4, we can see the mosaic plot and the BECCR Prediction Bubble Plot of the data, after considering the third variable G. Looking at these figures, we can see that both the mosaic plot and the BECCR Prediction Bubble Plot are able to capture the linear relationship between X and Y after accounting for G, with the BECCR Prediction Bubble Plots displaying this linear pattern in a much clearer fashion.

[1] We chose $B = 1000$ for the sake of efficiency and saving time because of additional simulations performed but not reported in this thesis.

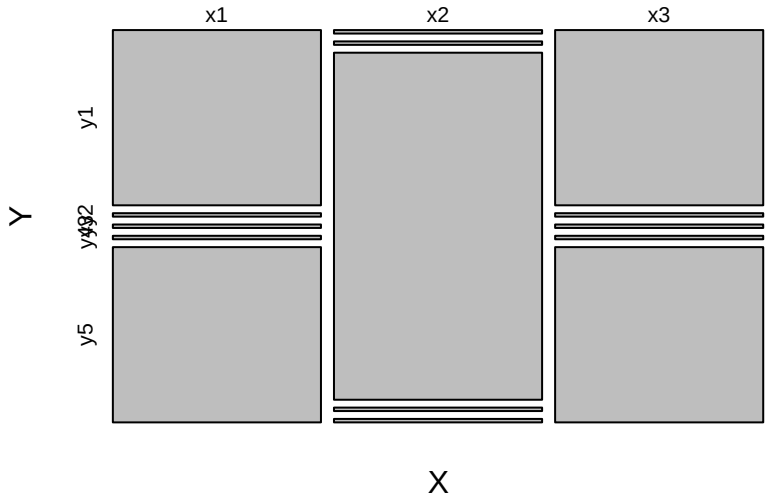

Figure 1: Mosaic Plot of Aggregated Hypothetical Data Set 1 without accounting for G

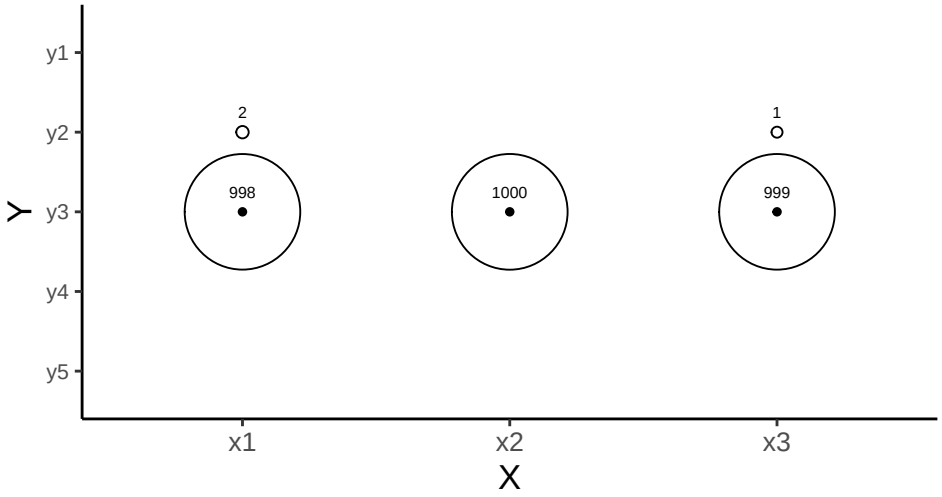

Figure 2: BECCR Prediction Bubble Plot of Aggregated Hypothetical Data Set 1 without accounting for G

Table 1: Hypothetical Ordinal Data with Linear Relationship Between X and Y

Y \ (G,X)	(g_1,x_1)	(g_1,x_2)	(g_1,x_3)	(g_2,x_1)	(g_2,x_2)	(g_2,x_3)
y_1	100	1	1	1	1	100
y_2	1	1	1	1	1	1
y_3	1	100	1	1	100	1
y_4	1	1	1	1	1	1
y_5	1	1	100	100	1	1

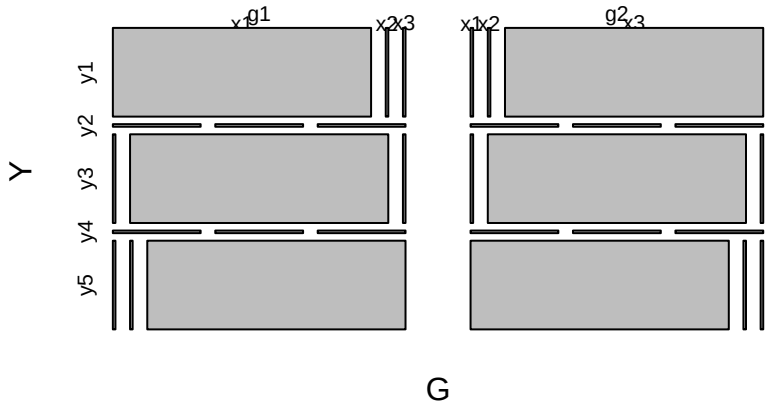

Figure 3: Mosaic Plot for Hypothetical Data with Linear Relationships after Accounting for G

```
createBubblePlot2(var_index = 1, regression_data = sim_predictions,
                  data = sim_data) +
  labs(
    title = ""
  )
```

Simulating Hypothetical Data with Quadratic Relationship

Moving onto the next case, we want to determine how well BECCR Prediction Bubble Plots do to visualize the following hypothetical health data oulined in Table 2. Without considering G, both the mosaic plot and the BECCR Prediction Bubble Plot (seen in Figures 5 and 6 respectively) seem to indicate no relationship between X and Y, which deviates what we observed from the data in Table 2. On the other hand, when looking at Figure 7 and Figure 8, one can see a quadratic relationship between X and Y within each level of the BECCR Prediction Bubble Plot, but not so easily in the mosaic plot.

It is important to notice that the BECCR Bubble Prediction Plot seems to "condense" the results, and this is due to how the (checkerboard copula) regression works, as well as the fact we are "predicting" a

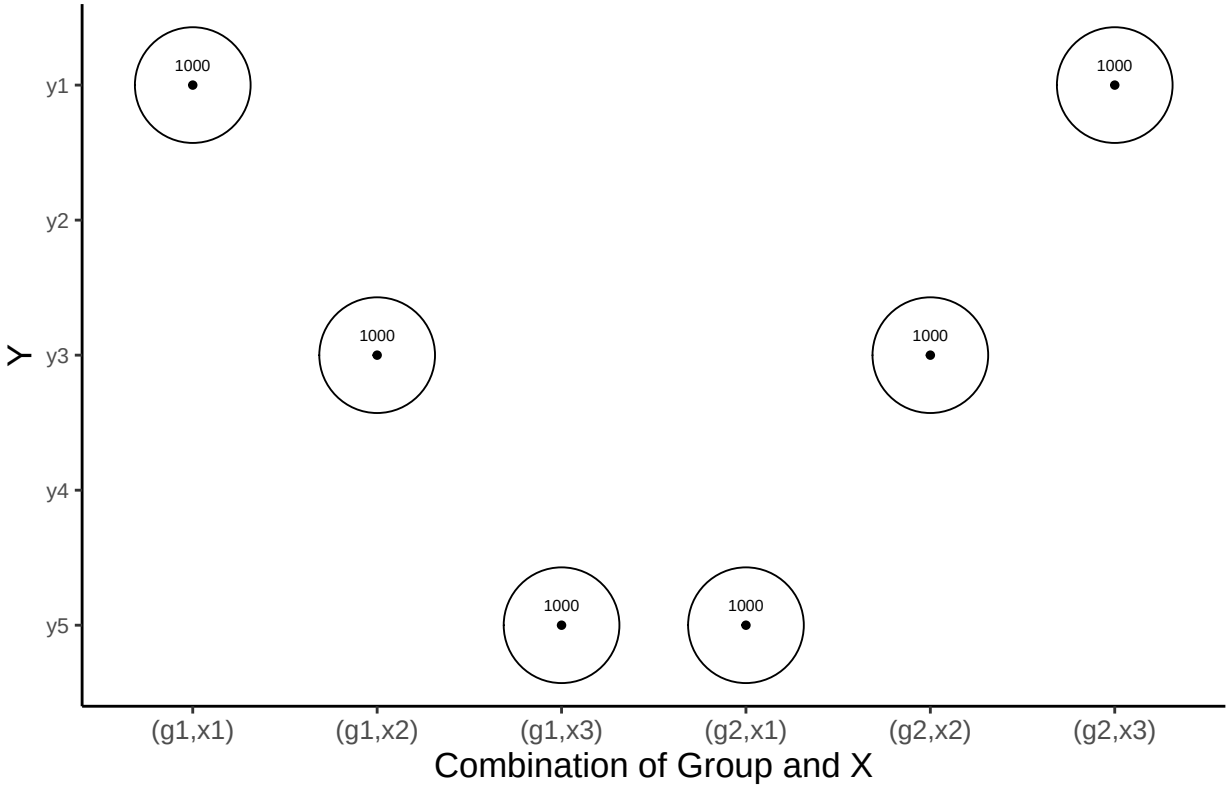

Figure 4: BECCR Prediction Bubble Plots for Hypothetical Data with Linear Relationships after Accounting for G

response category. As a result, the quadratic relationship of X and Y is better captured and displayed in the BECCR Prediction Bubble Plot (Figure 8), while some people might have difficulties visualizing this quadratic relationship in the mosaic plot (Figure 7).

Table 2: Hypothetical Ordinal Data with Quadratic Relationship Between X and Y

Y \ (G, X)	(g_1, x_1)	(g_1, x_2)	(g_1, x_3)	(g_2, x_1)	(g_2, x_2)	(g_2, x_3)
y_1	0	60	0	30	0	30
y_2	15	30	15	25	10	25
y_3	20	20	20	20	20	20
y_4	25	10	25	15	30	15
y_5	30	0	30	0	60	0

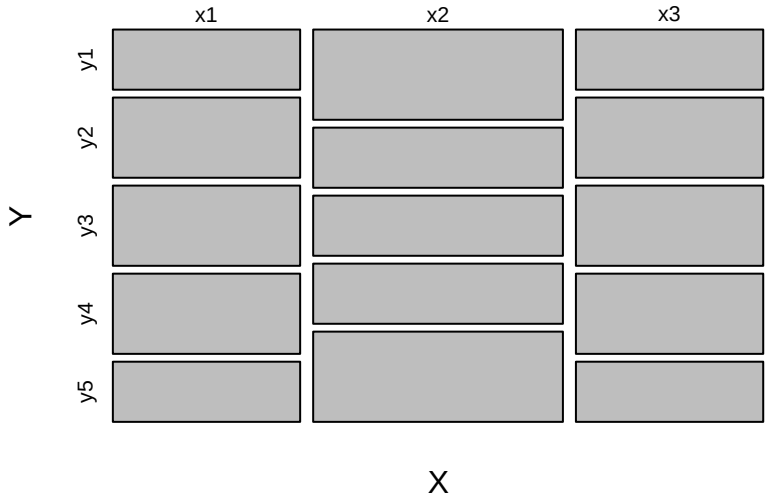

Figure 5: Mosaic Plot of Hypothetical Ordinal Data with Quadratic Relationship without Accounting for G

```
createBubblePlot2(var_index = 1, regression_data = sim_predictions,
                  data = sim_data) + labs(title = "")
```

Simulating Hypothetical Ordinal Data with a Pattern

For our third simulation, we want to see how well BECCR Prediction Bubble Plots can do to visualize data with different patterns between X and Y for different levels of G. The hypothetical data set for this case is outlined in Table 3, which shows an irregular relationship between X and Y for $G = g_1$, but there still seems to be a pattern that exists.

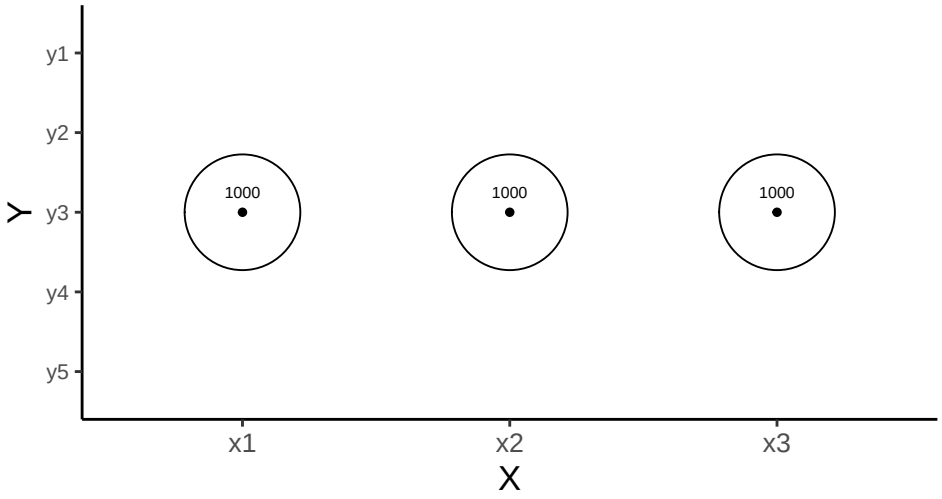

Figure 6: BECCR Prediction Bubble Plots of Hypothetical Ordinal Data with Quadratic Relationship without accounting for G

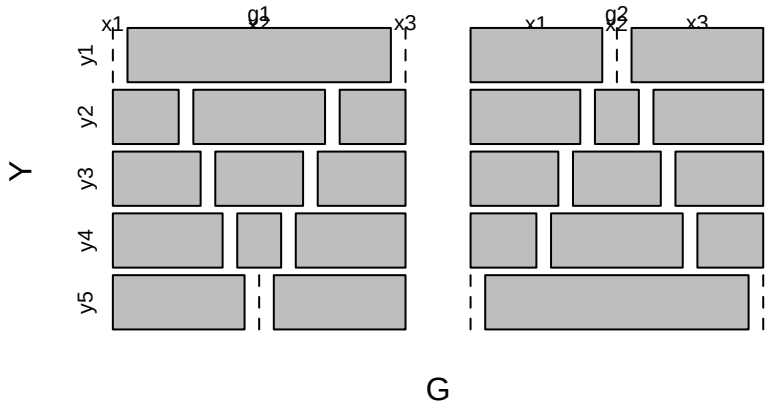

Figure 7: Mosaic Plot for Hypothetical Ordinal Data with Quadratic Relationship between X and Y after accounting for G

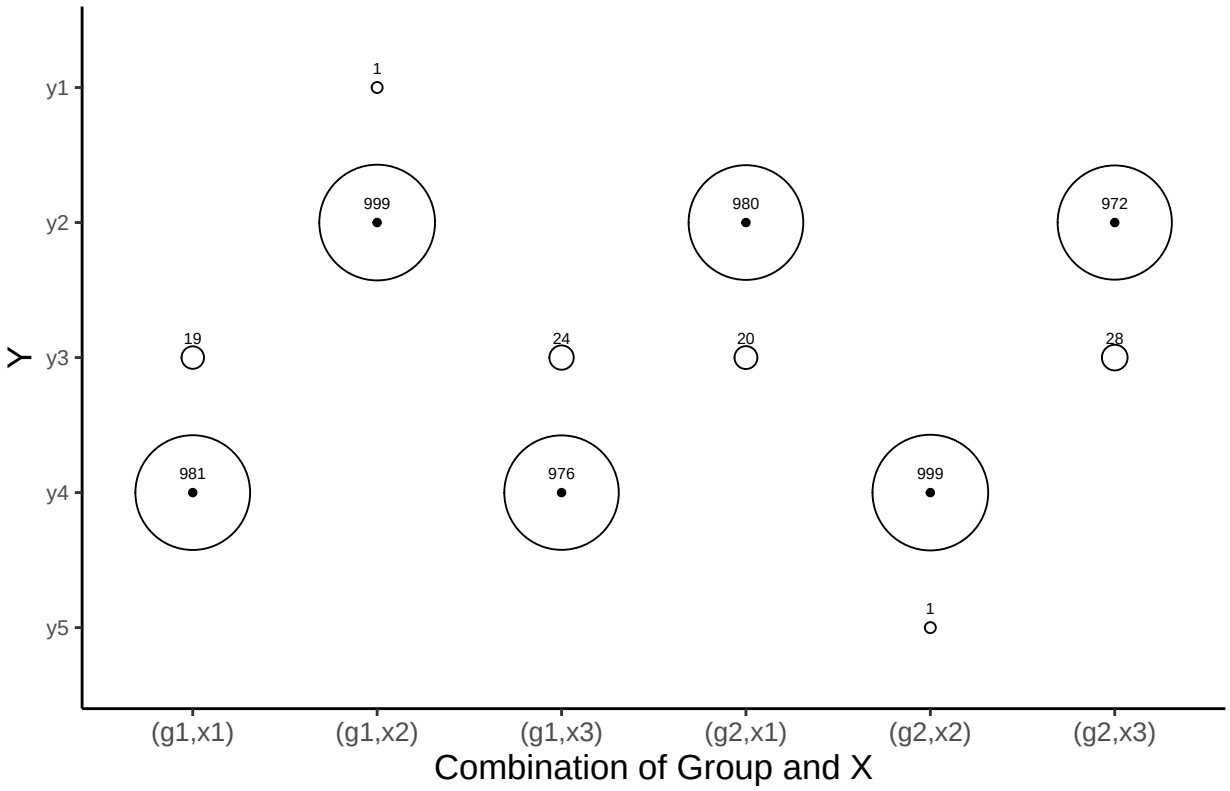

Figure 8: BECCR Prediction Bubble Plots for Hypothetical Ordinal Data with Quadratic Relationship after accounting for G

Again, when not considering G, the visualizations seem to indicate no relationship between X and Y for both the mosaic plot and the BECCR Prediction Bubble Plot (seen in Figures 9 and 10 respectively). It may be worth noting that in the BECCR Prediction Bubble Plot (in Figure 10), some aspects of the pattern relationship seems visible, and may suggest that BECCR Prediction Bubble Plots applied to aggregated data without the grouping variable G can still potentially offer some insights to the relationships of the variables.

Table 3: Hypothetical Ordinal Data with a Relationship Pattern Between X and Y

Y \ (G,X)	(g_1,x_1)	(g_1,x_2)	(g_1,x_3)	(g_2,x_1)	(g_2,x_2)	(g_2,x_3)
y_1	50	50	1	1	1	150
y_2	1	1	1	1	1	1
y_3	50	1	50	1	150	1
y_4	1	1	1	1	1	1
y_5	1	50	50	150	1	1

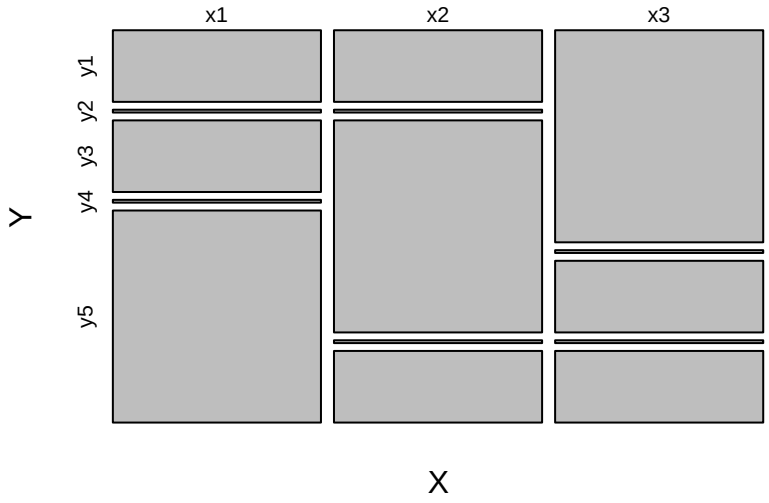

Figure 9: Mosaic Plot of Hypothetical Ordinal Data with a Pattern Relationship, aggregated without considering G.

Looking at the mosaic plot in Figure 11, we can clearly see that the mosaic plot is able to capture the different relationships between X and Y for $G = g_1$ for different levels of G. Similarly, the BECCR Prediction Bubble Plot in Figure 12 is also able to capture the relationship between X and Y, depending on G, again in a much cleaner manner. While both plots successfully capture the peculiar pattern of X and Y's relationship, we begin to see some differences between the two plots. In particular, the part of the mosaic plot for g_1 was shorter than the other part of the plot for g_2, which arises due to the different sizes of the two subgroups. In the previous simulations, each subgroup g_1 and g_2 have equal sizes, but when the proportions change, the

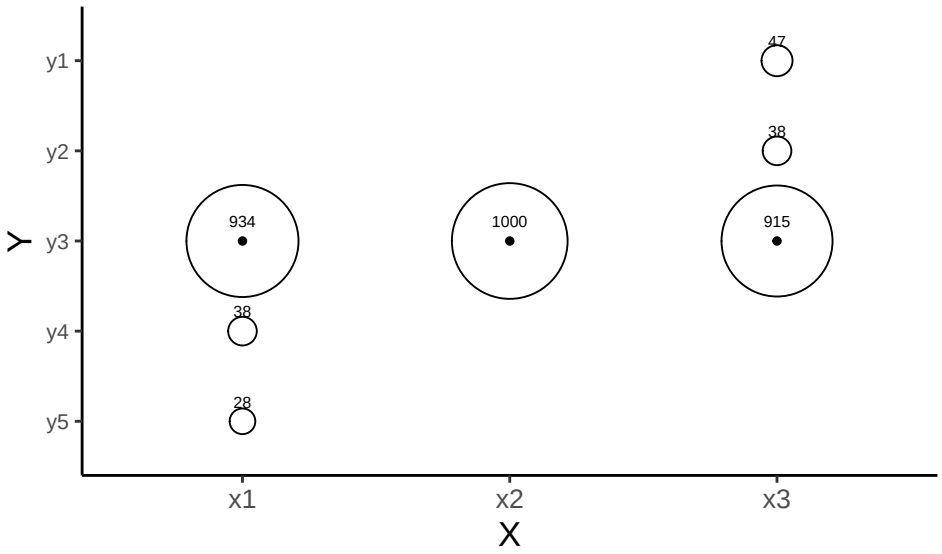

Figure 10: BECCR Prediction Bubble Plot of Hypothetical Ordinal Data with a Pattern Relationship, Aggregated without considering G

mosaic plots will change with it. While this does not count as a drawback of mosaic plots, this observation motivates us to simulate the next case.

```
createBubblePlot2(var_index = 1, regression_data = sim_predictions,
                  data = sim_data)
```

Simulating Hypothetical Ordinal Data with Subgroups of Differing Sizes

As mentioned, this simulation will be performed on a hypothetical data set that will have g_1 be in much smaller proportion than g_2, as we want to see how well BECCR Prediction Bubble Plots can help us visualize the interactions of it. This data set can be seen in Table 4. Clearly, the proportion of observations in g_1 is much lower than those of g_2. We can see this exhibited as the relationship in g_2 seems to dominate the visualizations in Figures 13 and 14.

Table 4: Hypothetical Ordinal Data with Differing Proportions of Subpopulations in g_1 and g_2

Y \ (G, X)	(g_1, x_1)	(g_1, x_2)	(g_1, x_3)	(g_2, x_1)	(g_2, x_2)	(g_2, x_3)
y_1	0	10	0	100	0	100
y_2	5	0	5	100	0	100
y_3	5	0	5	100	0	100
y_4	5	0	5	100	0	100
y_5	5	0	5	0	200	0

We now want to see how the visualizations of the data between the mosaic plot and the BECCR Prediction Bubble Plot changes after accounting for G. In Figure 15, we can clearly see that the mosaic plot does capture the quadratic relationship inherent in the data. However, the difference between the sizes of g_1 and g_2 is extremely prominent. If the relationship was not as easy to visualize, we can imagine that it

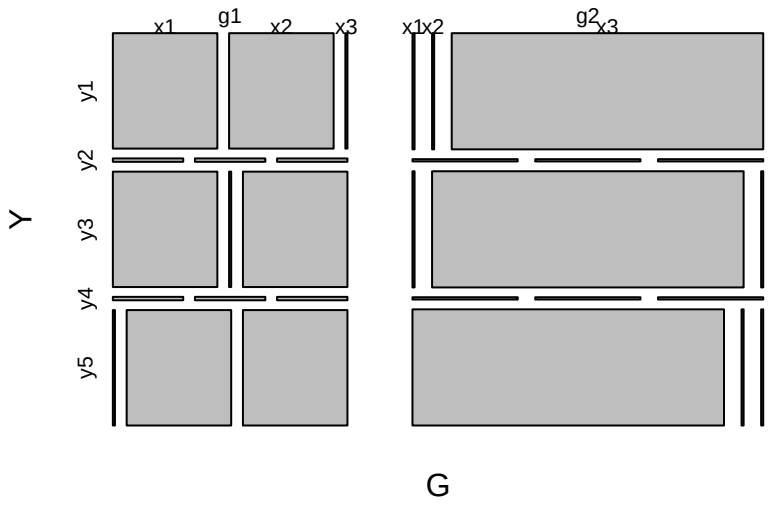

Figure 11: Mosaic Plot for Hypothetical Ordinal Data with Pattern After Accounting for G

will be hard to distinguish possible relationships of X and Y for $G = g_1$. On the other hand, the BECCR Prediction Bubble Plot in Figure 16 not only captures the quadratic relationship between X and Y, but also helps us visualize the different relationships between X and Y depending on what G is. While the data does get "condensed" again, the overall quadratic relationship is depicted much more clearly than it does in the mosaic plot in Figure 15. We expect that these differences will be further exacerbated in a high-dimensional example with more than 3 variables, as we will see in the next simulation case.

```
createBubblePlot2(var_index = 1, regression_data = sim_predictions,
                  data = sim_data)
```

Simulating Hypothetical High-Dimensional Ordinal Data

Table 5: Hypothetical Ordinal Data when $Z = z_1$,

Y \ (G, X)	(g_1, x_1)	(g_1, x_2)	(g_1, x_3)	(g_2, x_1)	(g_2, x_2)	(g_2, x_3)
y_1	0	10	0	150	0	150
y_2	5	0	5	150	0	150
y_3	5	0	5	150	0	150
y_4	5	0	5	150	0	150
y_5	5	0	5	0	250	0

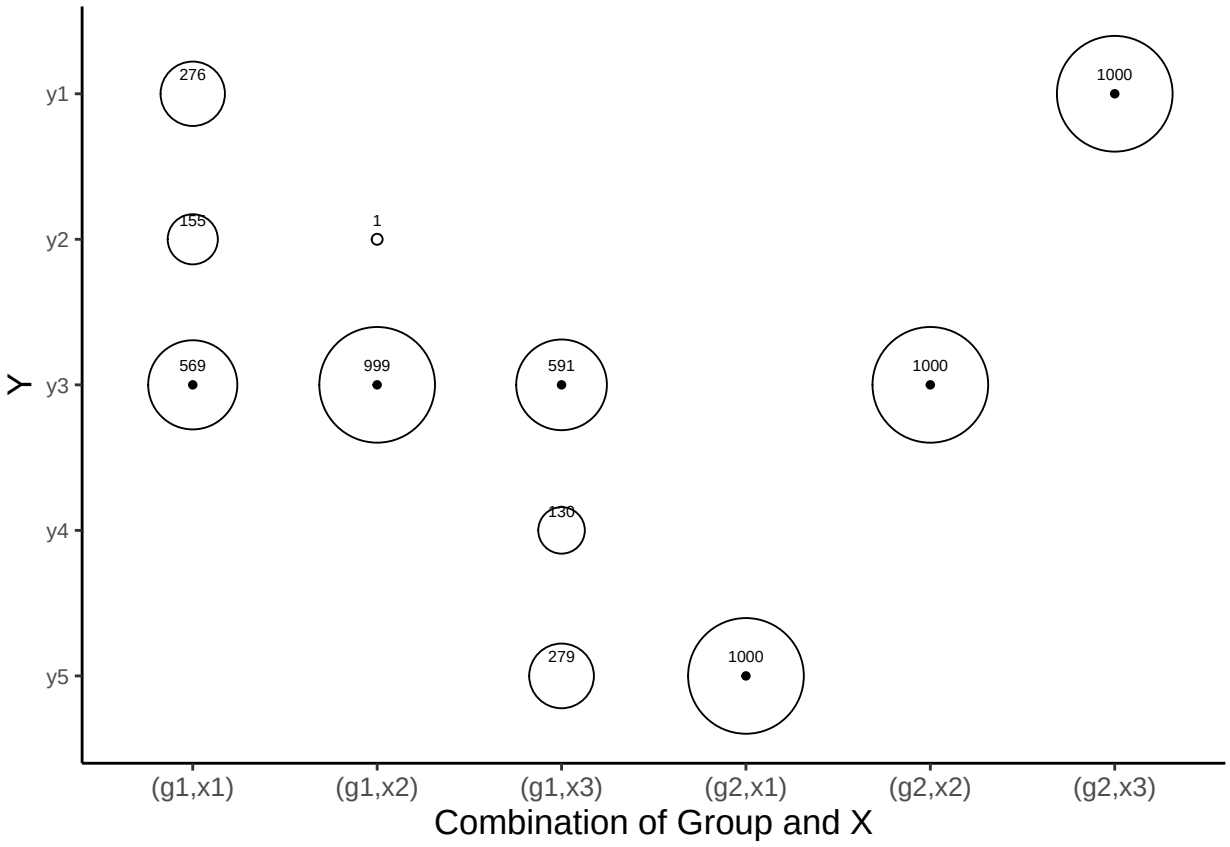

Figure 12: BECCR Prediction Bubble Plot of Hypothetical Ordinal Data with Pattern After Accounting for G

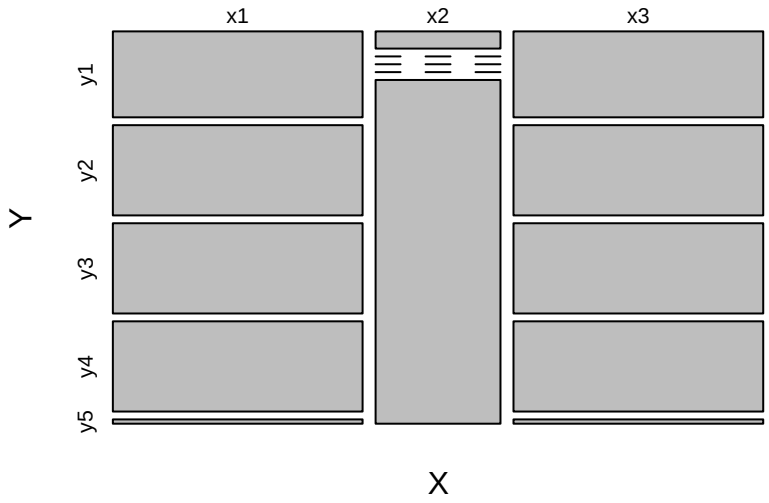

Figure 13: Mosaic Plot of Ordinal Data with Differing Proportions without accounting for G

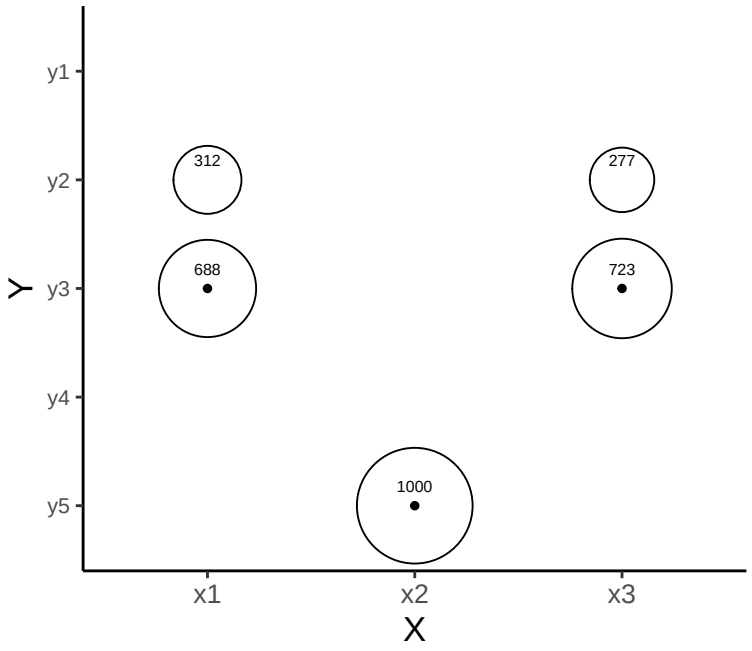

Figure 14: BECCR Prediction Bubble Plot of Ordinal Data with Differing Proportions without Accounting for G

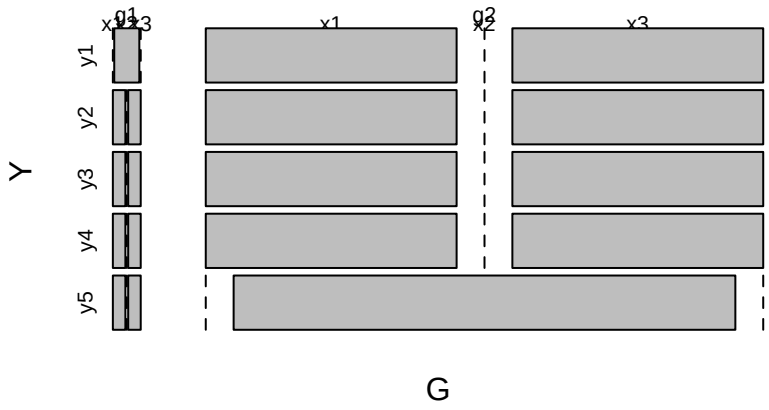

Figure 15: Mosaic Plot for Hypothetical Ordinal Data with Differing Proportions after Accounting for G.

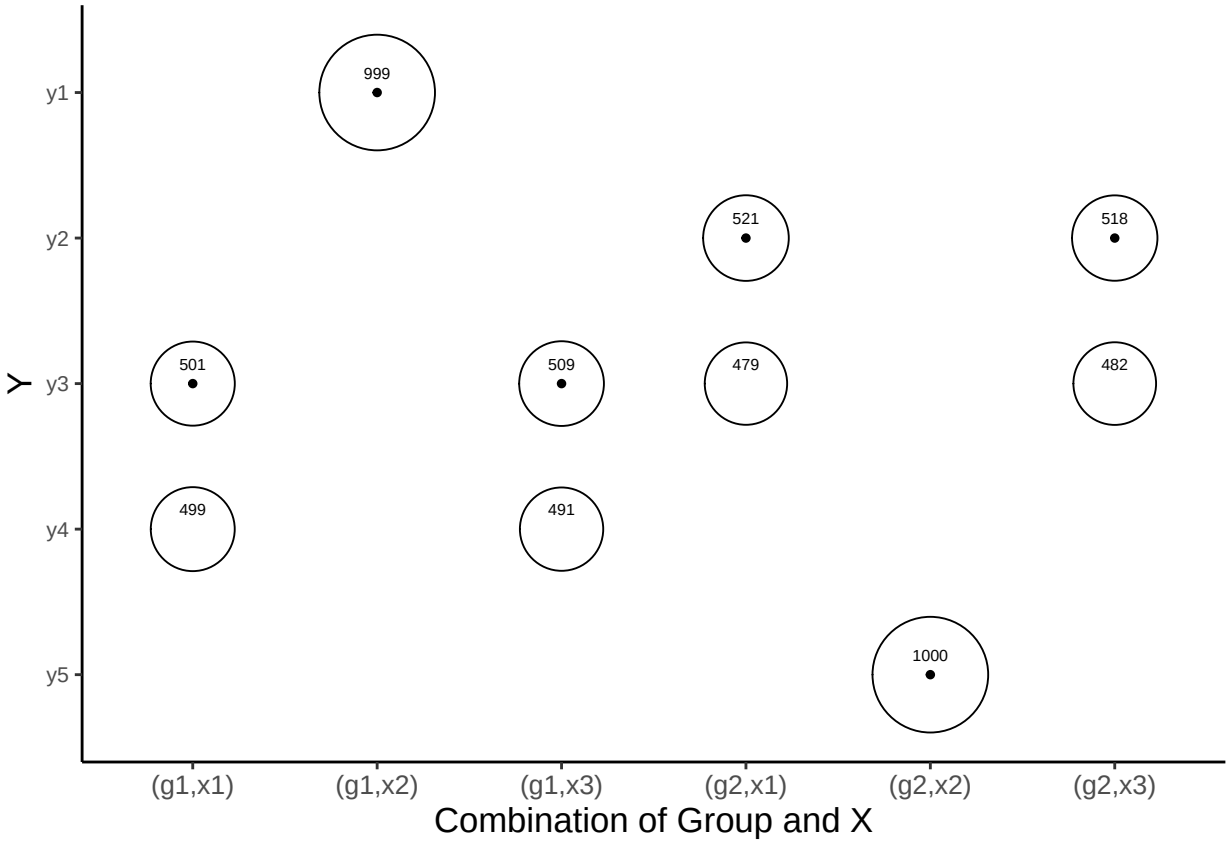

Figure 16: BECCR Prediction Bubble Plot of Hypothetical Ordinal Data with Differing Proportions after Accounting for G

Table 6: Hypothetical Ordinal Data when $Z = z_2$

Y \ (G, X)	(g_1, x_1)	(g_1, x_2)	(g_1, x_3)	(g_2, x_1)	(g_2, x_2)	(g_2, x_3)
y_1	150	1	1	3	3	10
y_2	1	1	1	3	3	3
y_3	1	150	1	3	10	3
y_4	1	1	1	3	3	3
y_5	1	1	150	10	3	3

For this simulation, we are going deal with a similar data set as in the prior simulations, but with an additional explanatory binary predictor $Z = \{z_1, z_2\}$. As we can see in Tables 5 and 6 (tables for $Z = z_1$ and $Z = z_2$ respectively), a combination of all the prior hypothetical data is exhibited. Without considering both G and Z the aggregated plots are shown in Figures 17 and 18 for the mosaic plot and the BECCR Prediction Bubble Plot respectively. From the BECCR Prediction Bubble Plot, we can clearly see that there seems to have a discernible association between Y and X, without accounting for the other variables, G and Z.

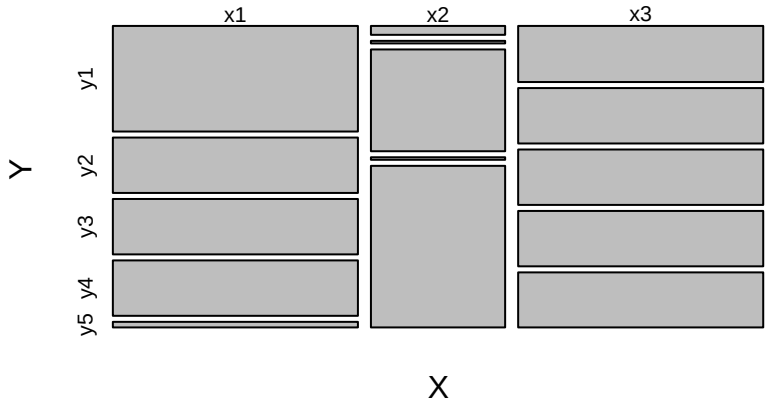

Figure 17: Mosaic Plot of Hypothetical High-Dimensional Aggregated Ordinal Data without accounting for G and Z.

```
createBubblePlot2(var_index = 1, regression_data = sim_predictions_ungroup_z,
                  data = sim_data_ungroup_z)
```

How would these plots change after adding G and Z to the checkerboard copula regression? We are particularly interested in how effective BECCR Prediction Bubble Plots could capture and identify the potential interactions between 3 predictors compared to the mosaic plots. These visualizations are shown in Figures 20 and 21. Notice that the top plot in Figure 21 is for $Z = z_1$, and the bottom plot is for $Z = z_2$. Immediately, we can see that the mosaic plot (Figure 20) is really convoluted to look at, difficult to interpret, and provides little useful insights about the dependence structure of the variables.

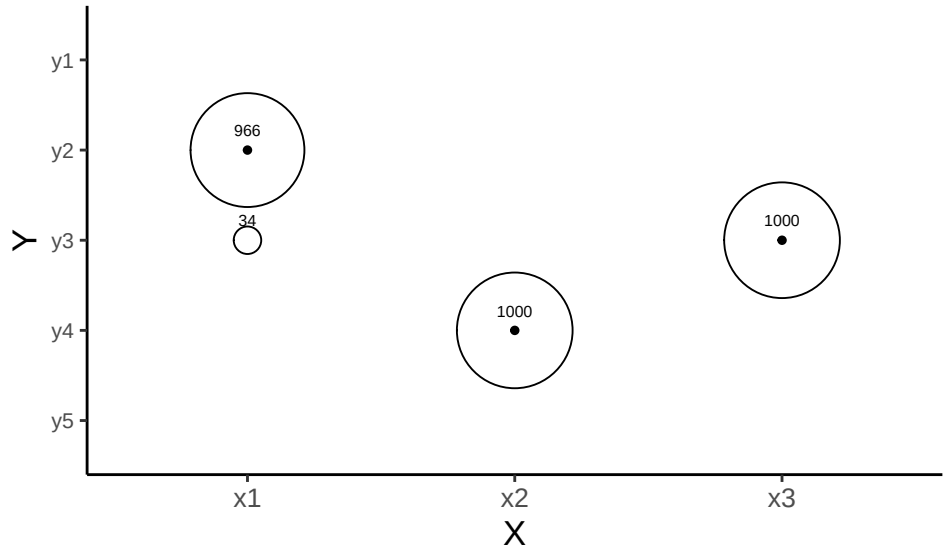

Figure 18: BECCR Prediction Bubble Plot of Hypothetical High-Dimensional Aggregated Ordinal Data without accounting for G and Z.

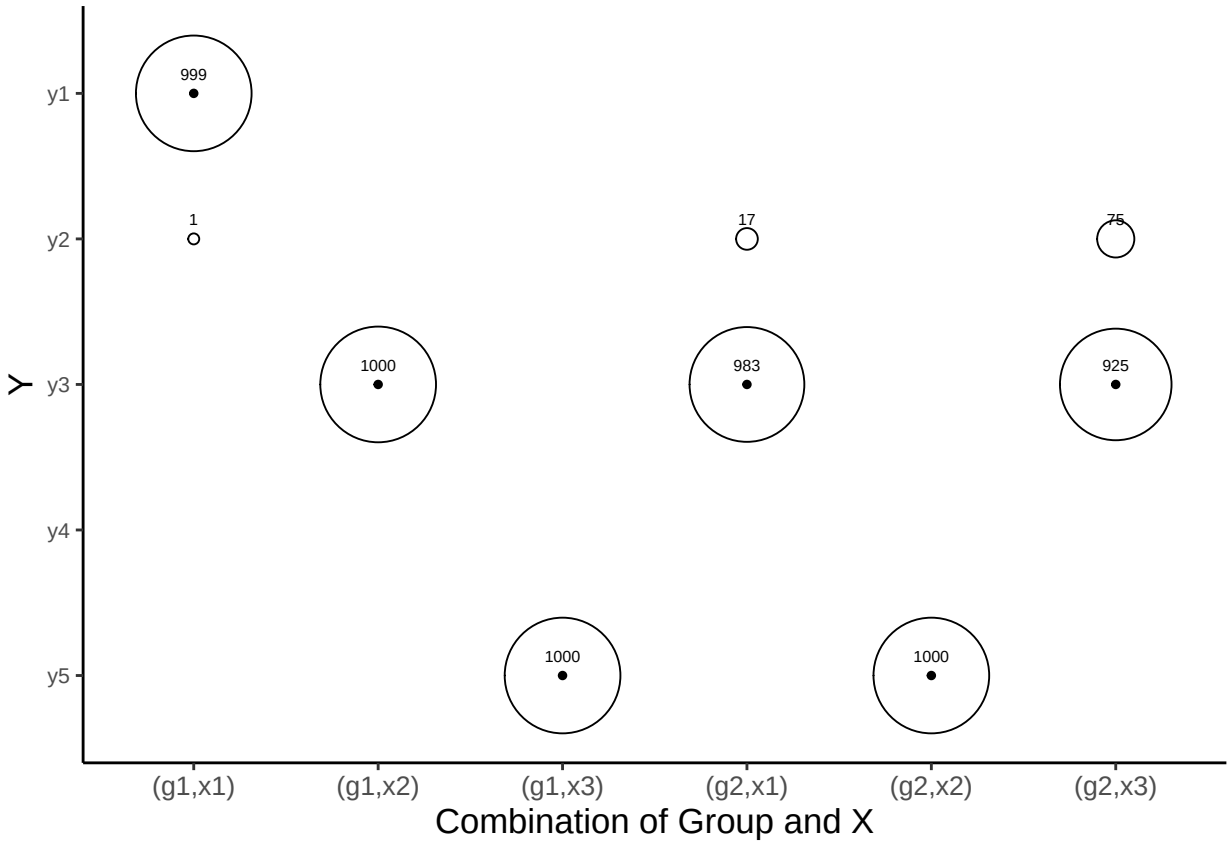

Figure 19: BECCR Prediction Bubble Plot for Hypothetical Multivaraite Ordinal Data without Accounting for Z

On the contrary, the BECCR Prediction Bubble Plot (Figure 21) clearly shows that for each combination of variables (Z, G, X), there is a noticeable relationship with Y. When $Z = z_1$, if $G = g_1$, there seems to be a "negative" quadratic relationship between X and Y; if $G = g_2$, there instead seems to be a positive quadratic relationship. Further, when $Z = z_2$, if $G = g_1$, there seems to be a strong negative linear relationship between X and Y; if $G = g_2$, there seems to be a moderate positive relationship between X and Y.

Comparing these results to the data in Table 5 and 6, we can see the BECCR Prediction Bubble Plot is able to convey these relationships clearly. In comparison, the mosaic plot is not only confusing to interpret when there are more than 3 variables, but also fails in identifying any meaningful dependence structure among these variables.

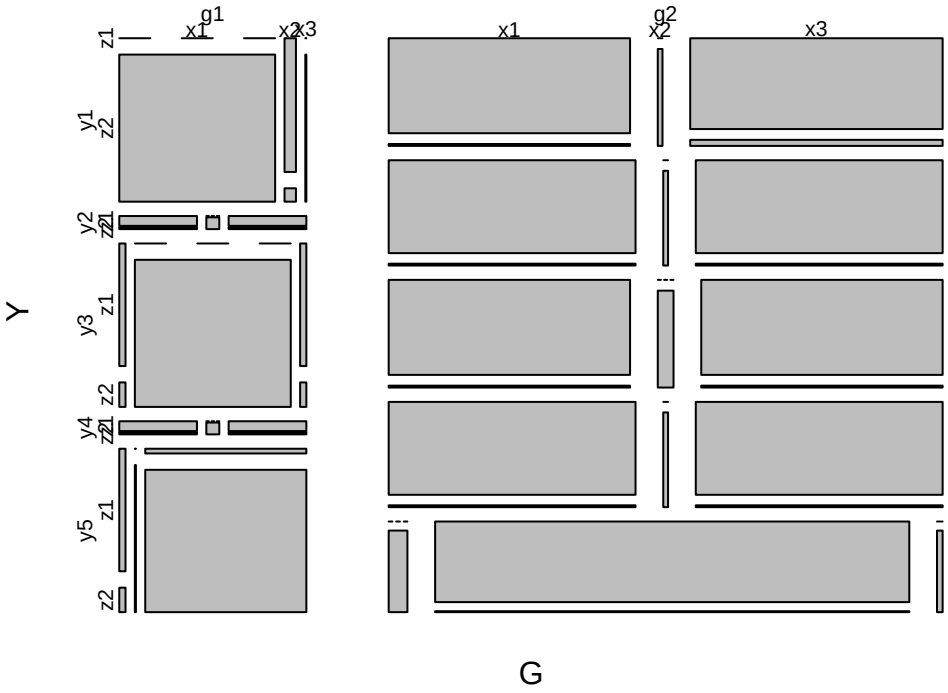

Figure 20: Mosaic Plot for Hypothetical Multivariate Ordinal Data after accounting for both G and Z.

Discussion

In summary, through the above simulations, we demonstrate how effective the proposed BECCR Prediction Bubble Plots can be used to capture and visualize various dependence structures — no matter whether they are linear, non-linear, different among different subgroups, and/or have subgroups of different sizes — in high-dimensional contingency tables. While one can use the usual mosaic plots to obtain similar insights about those dependence structures — when they are interpreted correctly and when the dimension of the data is no more than three, some might have a difficult time seeing those patterns easily and clearly, compared to what they see in the corresponding BECCR Prediction Plots. Not to mention that the BECCR Prediction

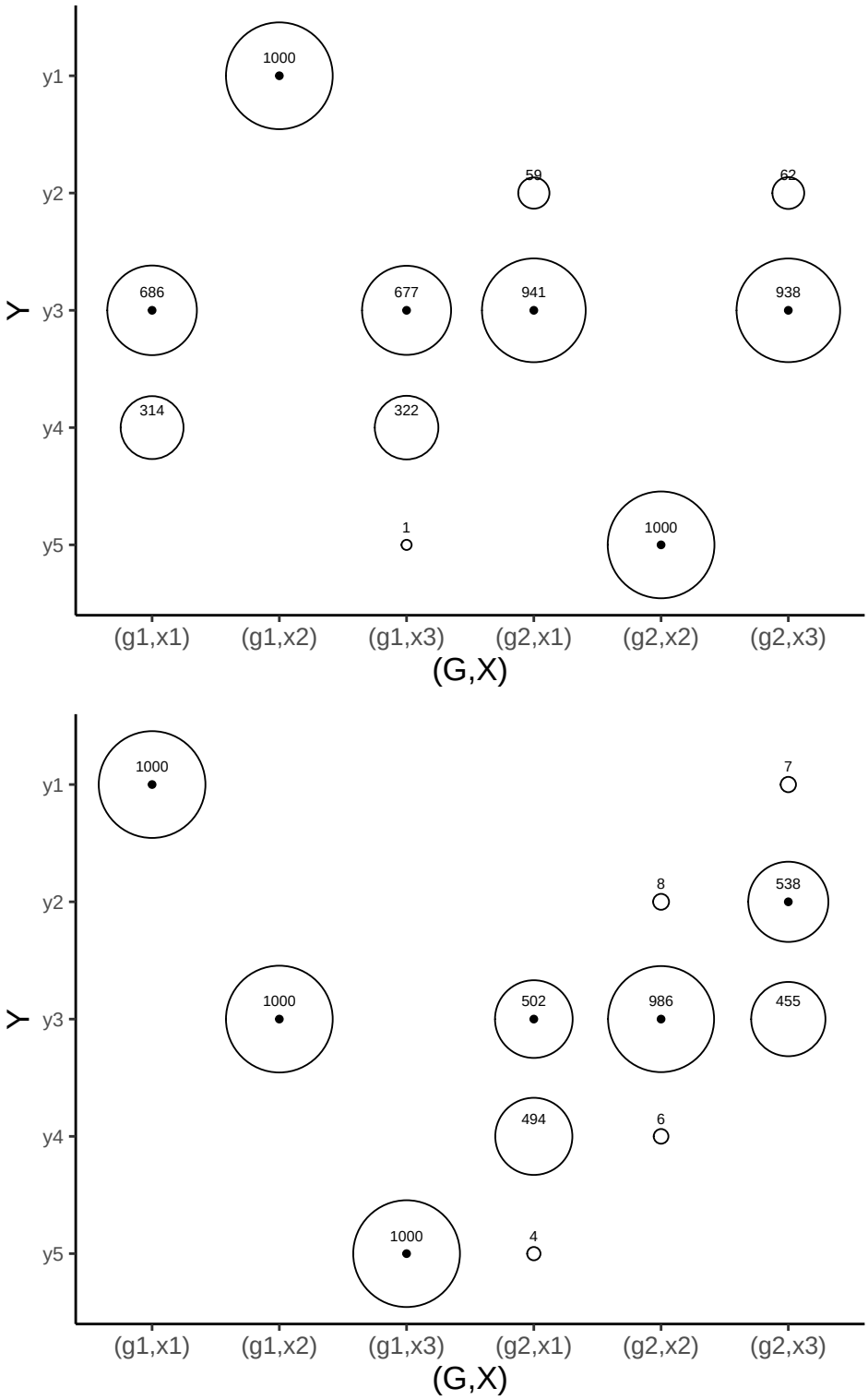

Figure 21: BECCR Prediction Bubble Plot for Hypothetical Multivariate Ordinal Data after accounting for both G and Z, with $Z = z_1$ being the top figure and $Z = z_2$ being the bottom figure.

Plots appear to have absolute advantages over mosaic plots when the dimension of the data is greater than three.

However, it's important to emphasize that our main goal here is not to show that BECCR Prediction Bubble Plots "outperform" the mosaic plots. Rather, we are proposing to use BECCR Prediction Bubble Plots as a *complementary* EDA tool and technique for visualizing dependence structures in categorical data, on top of mosaic plots and/or any existing visualization methods. One way to look at the mutual complement between BECCR Prediction Bubble Plots and mosaic plots for categorical data is to think of the role of Least-Squares (LS) fitted lines to scatter plots for quantitative data. No one would claim that LS-fitted lines are more important than scatter plots when visualizing relationships between quantitative variables; we use scatter plots to display observed data while using LS-fitted lines to capture potential *patterns* in quantitative data, to figure out what we should do (which models we may want to fit) at the modeling step of data analysis. Similarly, one can use mosaic plots to show the observed counts/proportions of the data, while further using BECCR Prediction Bubble Plots to help identify critical dependence structures and potential interactions in categorical data before moving to fit a model on those data. It's our firm belief that this newly developed graphical tool can help us do a better job in modeling categorical data.

Real World Data Application

Chronic pain in adults occur more and more frequently as we age, and this pain can often impact life or work activities. These kinds of high-impact chronic pains are one of the most common reasons why adults seek medical care, and the pain is associated with decreased quality of life, opioid dependence, and poor mental health. As such, it is important to study how different people are affected by pain, and what factors might be associated with one's perception of pain is important to study. As pain is a subjective measure, there could be tendencies within certain subpopulations who may categorize their pain differently, and we want to investigate what factors could impact this.

To accomplish this task, we used data from 2019 National Health Interview Survey (NHIS) from the National Center for Health Statistics in the CDC. We use the Sample Adult interview data[2] and will use the following variables adapted from this data set:

- Amount of Pain, response variable Y, which is an ordinal variable with 3 levels: Small Pain ($y_1 = \text{SP}$), Moderate Pain ($y_2 = \text{MP}$), and Great Pain ($y_3 = \text{GP}$)
- Frequency of Pain, explanatory variable X, which is an ordinal variable with 3 levels: Low (x_1), Medium ($x_2 = \text{Med}$), and High (x_3)
- Ethnicity, explanatory variable G, which is a binary ordinal variable: People of the Global Majority ($g_1 = \text{POGM}$) and White (g_2)
- Sex, explanatory variable Z, which is a binary ordinal variable: Men (z_1) and Women (z_2)

The actual data set is outlined in the Table 7.

[2]originally $n = 31997$ but after filtering out all missing data for our desired variables, we are left with an analytic sample $n = 19185$.

Table 7: Perception of Amount of Pain Suffered by Pain Frequency, Sex, and Ethnicity

Pain Amount (Y)	Pain Frequency (X)	Sex (Z)	Ethnicity (G)	
			POGM (g_1)	White (g_2)
Small Pain (SP)	Low	Men	911	2520
		Women	1126	2649
	Medium (Med)	Men	58	264
		Women	54	221
	High	Men	54	346
		Women	75	232
Moderate Pain (MP)	Low	Men	481	1082
		Women	720	1427
	Medium (Med)	Men	120	441
		Women	212	663
	High	Men	170	708
		Women	231	859
Great Pain (GP)	Low	Men	134	293
		Women	266	440
	Medium (Med)	Men	69	132
		Women	125	236
	High	Men	189	523
		Women	337	817

Given this data set, we want to explore if it has any peculiar features and if we can glean any insights about the relationships about the variables. For each variable in the data, the table consisting of the proportion of observations in each category is listed in the Appendix .

Besides using those tables as a numerical summary of the data, we also want to utilize graphs to visualize as part of our EDA of the data set. The first visualization tool we use is the mosaic plot, shown as Figure 22.

From this mosaic plot, we can see that for every response of pain level, there doesn't seem to be much of a difference in the proportions of men and women responding based off their ethnicity. That is, the proportions of men and women for POGM who respond to each pain level seem to be proportionally similar to those of white men and women. Further, the relative differences in the proportions between the pain frequencies also seem similar between POGM and white participants.

However, it is unclear whether or not there are tangible interactions between the variables and what their relationship would be. To hopefully gain more useful insight into this, we create a BECCR Prediction Bubble Plot shown in Figure 23 to help.

Looking at Figure 23, one can easily see some clear relationships arise from the BECCR predicted categories of the amount of pain. For both POGM and White Men, they seem to share a similar clear positive relationship between pain frequency (X) and amount of pain perceived (Y). This relationship seems to be quite strong as the prediction jumps from Small Pain (SP) to Great Pain (GP) after increasing from low pain frequency to medium pain frequency. However, the results for Women tell a different story, depending on whether they are POGM or White. Those different stories imply that there is a potential interaction effect between Ethnicity (G) and Pain Frequency (X).

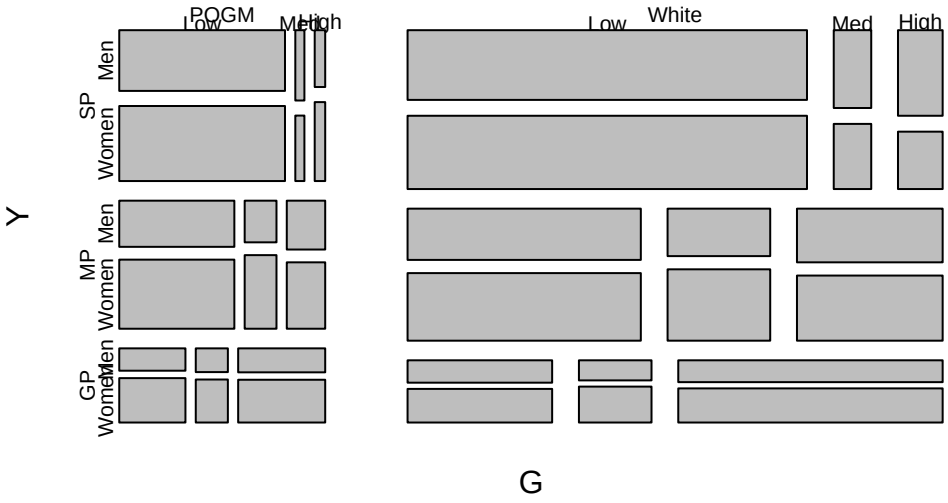

Figure 22: Mosaic Plot of Perception of Pain Adults Feel by Pain Frequency, Sex, and Ethnicity.

Similar to White Men, White Women also seem to have a positive relationship between Pain Frequency (X) and Pain Percieved (Y), but this relationship seems more linear than White Men's as we predict that White Women feel moderate pain (MP) when they have medium (med) pain frequency, instead of jumping directly up to great pain (GP) like it was for the White Men. On the other hand, for POGM Women, the amount of pain seems to be consistently high, even when the pain frequency is low. Though it dips back to moderate pain (MP) when the pain frequency is medium (med), it goes back to great pain (GP) when the pain frequency is high. This relationship is unlike the relationships we observe for POGM Men, as this seems quadratic in a sense. All those differences tell us that Sex (Z) and Ethnicity (G) might interact with each other.

Not only does this suggest that there may be differences in how one should interpret the pain data for men and women, but we also have to take into consideration their ethnicity as well; in other words, it's important to pay attention to the intersectionality between Sex (Z) and Ethnicity (G) when predicting a patient's pain level by their pain frequency (X).

To illustrate how ignoring intersectionality may lead Simpson's paradox, let us first consider the BECCR Prediction Bubble Plot without considering their Ethnicity (Z). This is illustrated in Figure 24. Taking a look at this plot, we can see that the pattern between the reported pain amount and reported pain frequency is the same strong positive relationship that we saw in Figure 23 for both White and POGM men. When you look at women without considering their ethnicity, we can see that the aggregated plot (the right half of Figure24) looks somehow similar to the BECCR Prediction Bubble Plot of POGM women (the right half of the top plot in Figure 23), with the White Women's positive linear relationship ("the right half of the bottom plot in Figure 23) being hidden. This is an effect of amalgamating our subpopulations together, and is what we are calling"Simpson's Paradox" for response variables with more than 2 categories in this paper. As a consequence, one may over-estimate White Women's pain level when they report low pain frequency and prescribe too strong pain medicine for those White Women.

We can take this another step further and consider the BECCR Prediction Bubble Plot of the NHIS data

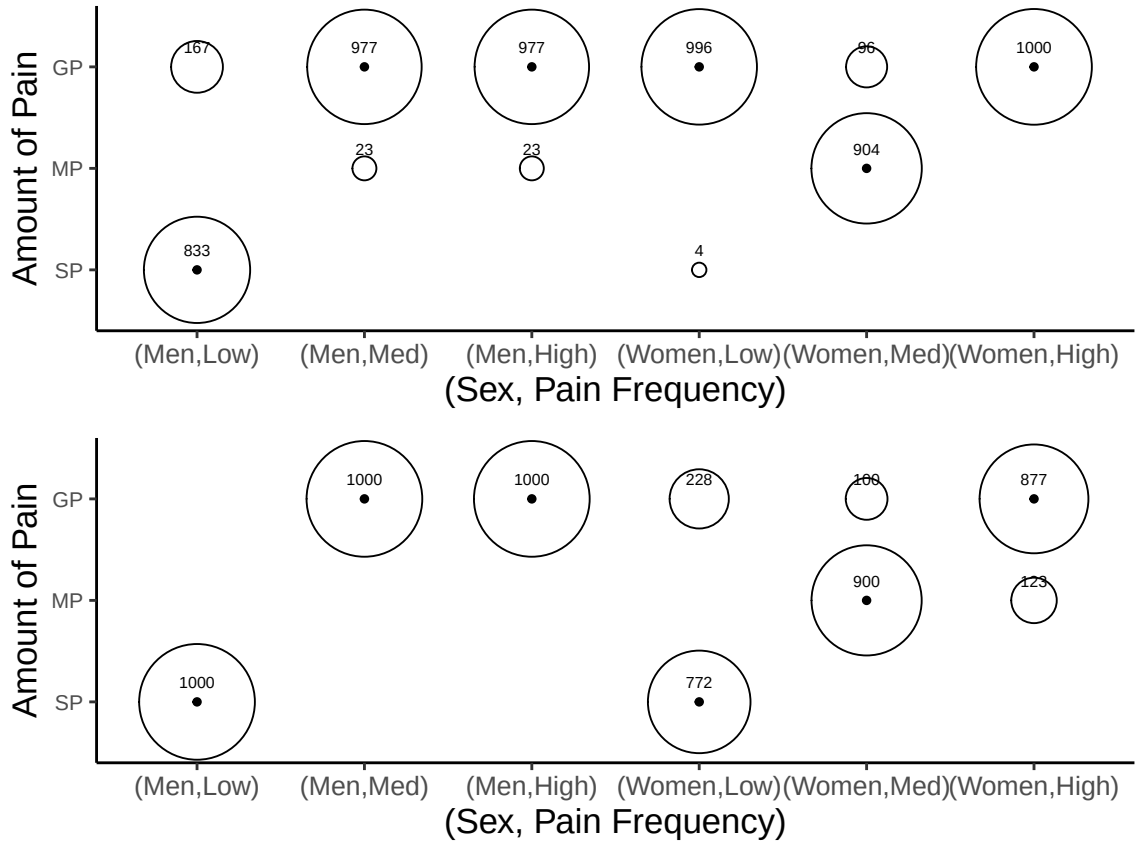

Figure 23: BECCR Prediction Bubble Plots of 2019 NHIS Adult Survey Data about Amount of Pain received by Reported Pain Frequency, Sex, and Ethnicity. (Top) Combination of categories (POGM, Sex, Pain Frequency) and (Bottom) combination of categories (White, Sex, Pain Frequency) estimated by copula regression in the 1000 bootstrap resampling

without considering both Ethnicity (G) *and* Sex (Z), which results in Figure 25.

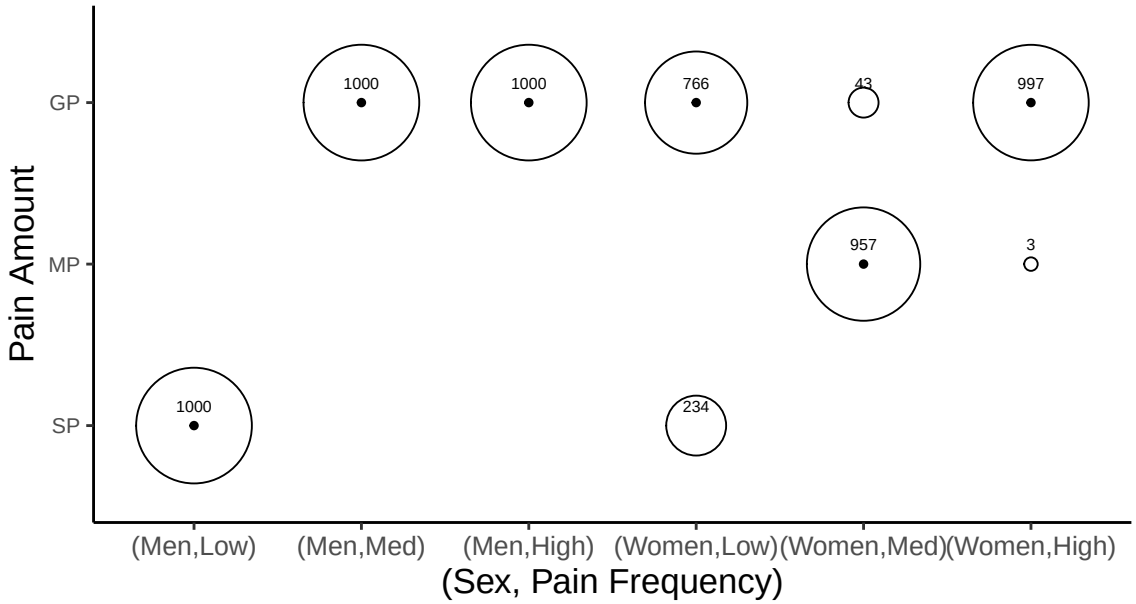

Figure 24: BECCR Prediction Bubble Plots of 2019 NHIS Adult Survey Data about Amount of Pain received by Reported Pain Frequency and Sex. Predicted category estimated by copula regression in the 1000 bootstrap resamples.

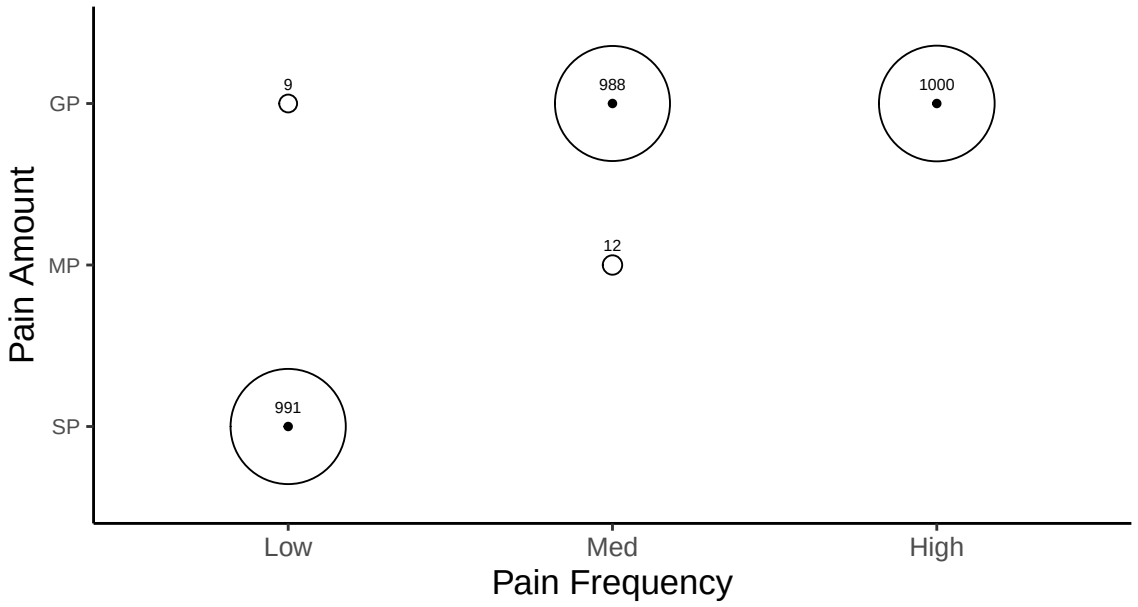

Figure 25: BECCR Prediction Bubble Plots of 2019 NHIS Adult Survey Data about Amount of Pain received by Reported Pain Frequency. Predicted category estimated by copula regression in the 1000 bootstrap resamples.

Appendix

EDA for Single Variables for NHIS Adult Pain Data

```
## Ethnicity
tally(~G, data = pain_data2, format = "proportion")
```

```
## G
##      POGM     White
## 0.2779255 0.7220745
```

```
## Sex
tally(~Z, data = pain_data2, format = "proportion")
```

```
## Z
##       Men     Women
## 0.4427938 0.5572062
```

```
## Frequency of Pain
tally(~X, data = pain_data2, format = "proportion")
```

```
## X
##       Low       Med      High
## 0.6280427 0.1352619 0.2366953
```

```r
## Amount of Pain
tally(~Y, data = pain_data2, format = "proportion")
```

```
## Y
##        SP        MP        GP
## 0.4435757 0.3708105 0.1856138
```