A Multivariate Approach to Ordinal Data

Analyzing emotional responses to abortion

Jackson Curtis

1 Introduction

Survey data is frequently multivariate in nature. Rarely do surveys contain just a single response in which we are interested in making inferences. Instead, they include many questions that may have small differences between them. While separate regression models may be appropriate when the responses are conditionally independent (given the covariates), questions that are naturally highly related should be modeled jointly in order to make the most precise inferences.

Another aspect of survey modeling we will address in this paper is the way in which respondents give their answer. Many surveys use Likert-scaled data to quantify responses on a spectrum. These responses generally make it easy for the respondents (as opposed to having them assign a number), but some information is undoubtedly lost. For example, most respondents could likely tell you whether their second choice would be the lower level or the higher level than the one they actually chose, but the surveys do not capture this. Additionally, the wording of the scale may favor one position as overly broad, so it is selected by many people even if they hold dissimilar views. Frequently these responses are assigned a numerical value and modeled using a normal distribution (or only the mean is estimated), but this approach ignores the small details. A latent variable ordinal model models the data under a more sound theory of how the respondents are actually answering the questions.

In this paper we will apply a multivariate ordinal model to four responses measuring the amount of compassion felt towards individuals who chose to have an abortion. Various explanatory variable will be examined. We will analyze the factors that contribute to compassion as well as the relationships between the four situations. Abortion is an emotionally charged issues, and understanding the views of the general public (especially regarding situational intricacies) is a top priority of social science researchers. This type of analysis will allow us to do a much deeper dive on exploring the complex range of emotions felt in different situations.

2 Data

The data was gathered by a Human Development major at BYU, in order to study the charged emotions surrounding abortion. The data was a convenience sample, administered using Qualtrics software and shared on Facebook. As such, we do not expect it to be strongly representative of any general population. For example, of the 207 respondents, 180 were female. The survey gathered basic demographic information (age, sex, marital status, income), a variety of measures of self-reported religiosity and political activity, and an assortment of measurements on the strength of emotional responses to abortion. The survey posed four situations that led to an abortion and asked the respondent to give a Likert-scale rating (ranging from 'Not at all' to 'Extreme') regarding how much they felt of each of the following emotions: compassionate, angered, disturbed, grieved, frustrated, impassioned, warm, tender, and alarmed. The four situations were:

Elective. "Liz was 23 years old when she became pregnant with her boyfriend. This was unplanned and they did not have future plans to get married. They both just started their first year in graduate school. Having a baby did not fit her career and educational goals. She wasn't planning on settling down until her late twenties. Liz and her boyfriend chose to have an abortion."

Rape. "April was homeless, 18 years old, and pregnant. Her pregnancy was the result of a rape. She has no education beyond junior high and extremely limited financial resources and family support. April chose to have an abortion."

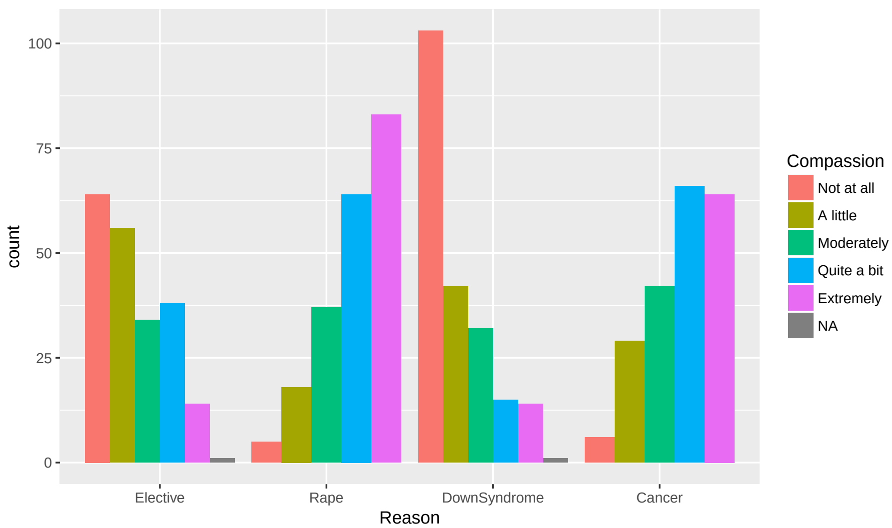

Figure 1: Compassion ratings for each of the four scenarios

Down Syndrome. "Britney, who is pregnant, and her husband James find out that their baby will have down syndrome. They chose to have an abortion."

Cancer. "Amanda, who is married, just found out that she is pregnant. Shortly after, she learned that she has cervical cancer. Her doctors told her that she wouldn't be able to receive treatment while she was pregnant, so she chose to have an abortion."

We will take interest in the compassion response to each of the four situations. Figure 1 shows the raw numbers for each of the four scenarios. People were likely to be strongly compassionate in the cases of rape and cancer, but much less compassionate in elective abortions and in cases of down syndrome.

Our primary interest lies in the relationships between the four scenarios. Figure 2 shows the raw Spearman correlation matrix for the four situations. Spearman is the preferred method for correlations of ordinal variables because it considers rank ordering and not magnitude. We see that all responses have a positive correlation, with rape and cancer being the most strongly related and rape and down syndrome being the least correlated. In our analysis we will answer the questions "What relationship do the covariates have to the four different responses?" and "How do the four scenarios relate after accounting for our covariates?"

We will briefly describe the covariates that we will investigate. Religious importance was a Likert-valued response to the question "How important is religion in your life?" Similarly, politically active was a response to the question "How politically active are you?" Income was a household income, broken into five bins, which we will treat as a continuous variable. Abortion active is a variable constructed from several questions indicating your willingness to take part in the abortion debate (whether posting online, protesting, etc.). Sex, married, and children were all binary variables. We plot age against a numeric version of the response with Loess smoothers in Figure 3 to get a feel for how best to include age in the model.

3 Statistical Model

We fit a latent variable ordinal model with the probit link using the mvord package in R. We will define our response as Y_{ij} where i is one of the 207 respondents and j is one of the four possible scenarios. Then:

$$Y_{ij} = r_j \tag{1}$$

r_j is the Likert response to situation j. Y_{ij} is assumed to be a discretized version of $\tilde{Y}_{ij}$ where $Y_{ij} = r_j$ if:

$$\theta_{j,r_{ij}-1} < \tilde{Y}_{ij} \leq \theta_{j,r_{ij}} \tag{2}$$

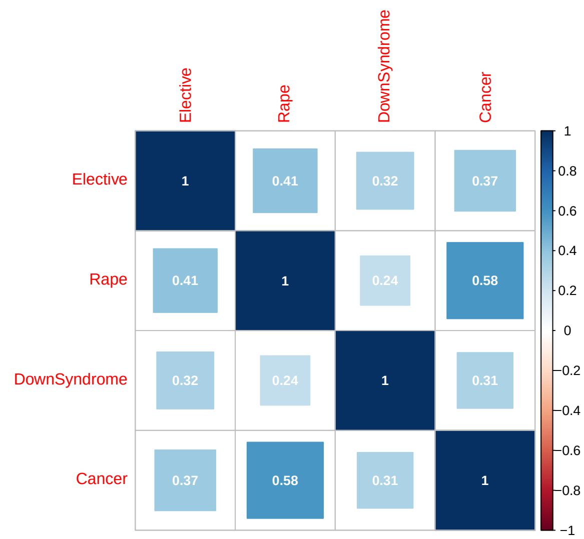

Figure 2: Spearman correlation matrix for responses

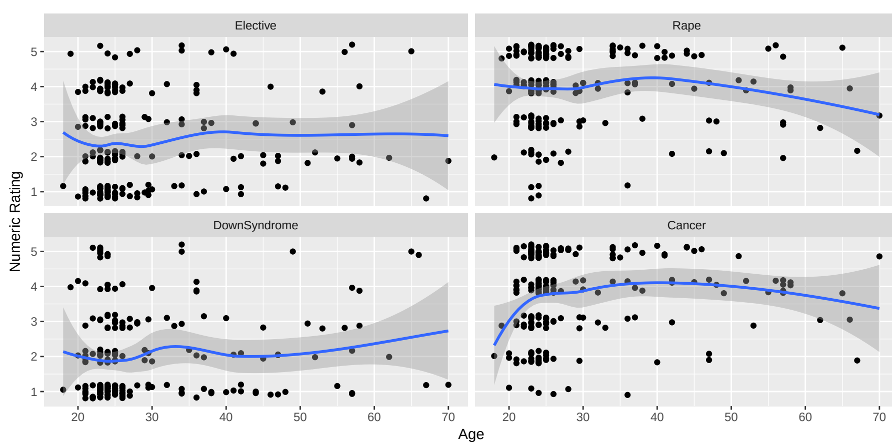

Figure 3: The response (jittered) plotted against age

$\tilde{Y}_{ij}$ is the latent variable which we will model as:

$$\tilde{Y}_{ij} = X\beta_j + \epsilon_{ij} \tag{3}$$

where β_j will be a unique coefficient vector for each scenario. ϵ_{ij} will be a single element of an error vector modeled as:

$$\epsilon_i \sim MVN(0, \Sigma) \tag{4}$$

Σ will be an unconstrained variance-covariance matrix determining the relationship between the errors on the latent variable.

Fitting this model involves estimating (in our case) four coefficients for each of our covariates (one for each scenario), four θ cutoffs that will determine the boundaries between our five Likert responses (for identifiability we fix two of these to be 0 and 1), and the four variance and six covariance terms that make up Σ.

Assumptions. This model assumes a certain psychology when answering these survey questions. It assumes that the respondent evaluates each question on a continuous spectrum, chooses where they're likely to be on that spectrum, and then identifies the Likert response that is closest to their position. One way in which you could see violation of this assumption is if you asked the respondents to choose their next preferred response after the one they chose. If they did not consistently choose one either just above or just below the first one they chose, the survey designer should either reconsider how the question was worded or consider switching to a nominal model.

The latent variable model also implies proportional odds. Since our coefficients only shift the latent variable higher or lower, it cannot simultaneously increase the probability of two categories at different ends of the spectrum. In general this seems like a safe assumption, but may not be ideal. For example, with our data we may hypothesize that a high score on the abortion activity covariate (an indicator of your willingness to engage in abortion debates) is likely to increase your polarization (you're probability of feeling either not at all or extremely compassionate) and decrease the probability of the moderate categories. Our model would be unable to catch this effect. We can pseudo-test whether this is a concern in this case by fitting univariate adjacent categories and testing the goodness of fit on the proportional odds model with the non-proportional odds model. The adjacent category non-proportional odds model could increase the probability of 'not at all' relative to 'a little' and the probability of 'extremely' relative to 'quite a bit,' whereas the proportional odds could only increase the probability of the higher or lower Likert value, not both. Running the χ^2 difference in deviance test for all four situations yields p-values which are all greater than 0.1, so we do not see clear evidence that our hypothesis is true or that the latent variable assumption is violated.

Model Selection. Model comparison will be done using Composite Likelihood AIC, which is built into the mvord package. Results will be shown for the full model using all covariates as well as for the model that minimized the AIC after exploring all main effects, reasonable interaction terms, and some non-linear transformations. Covariates can also be fixed to 0 for some scenarios but allowed to be estimated in others. However, because this greatly enlarges the number of possibilities, we will not explore these partially constrained models. We are mainly interested in assessing the relationship between the covariates and the responses, so we prefer the full model with all terms. However, we recognize that because most of the terms are insignificant (and the model has many coefficients), attempts to use that model to predict would introduce a lot of added variability and would result in much higher error rates than the reduced AIC model. Therefore we will assess the effects of the full model, but also report the findings of the model we would use if we were going to attempt to predict users responses from their covariates.

4 Model Results

The final reduced model that minimized AIC is:

$$\tilde{Y}_{ij} = \beta_{0j} + \log(\text{Age}) * \beta_{1j} + \text{Children} * \beta_{2j} + \text{ReligionImportance} * \beta_{3j} + \epsilon_{ij} \tag{5}$$

All the interactions between these variables increased AIC, as did adding any other covariates. The full model includes all covariates centered and scaled, but otherwise untransformed. Because more than 70% of our respondents were under 30, and it exhibited a long right tail, we experimented with fitting separate

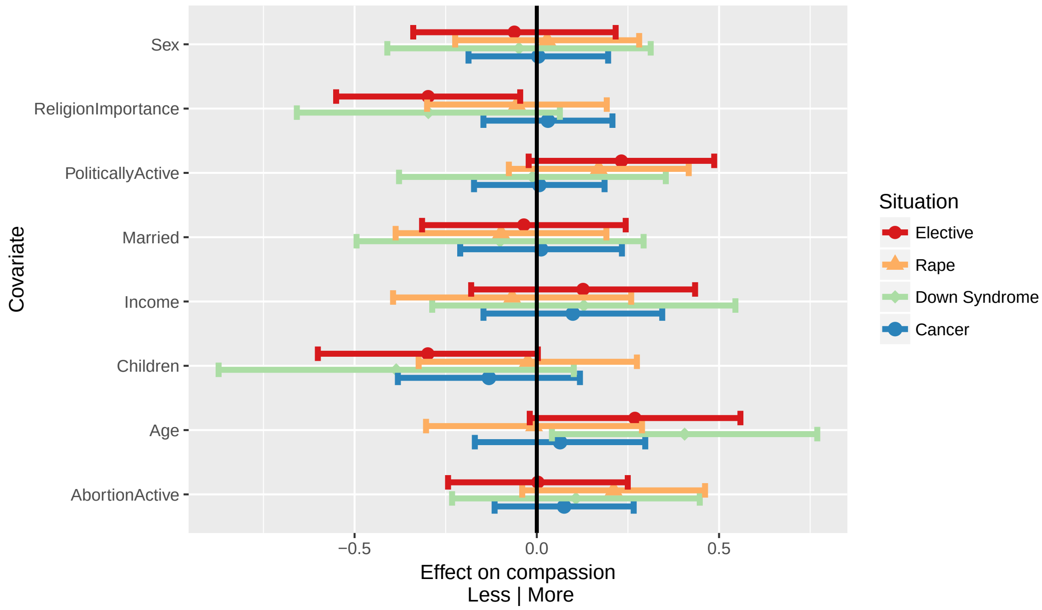

Figure 4: Point Estimate and 95% CI for coefficient for scaled covariate

effects for those under 30 and those above 30, but found the result to have worse AIC. Taking the log of age resulted in a better fitting model than untransformed age in both the full model and the reduced model, but the difference was so slight that we decided to maximize interpretability by leaving it unlogged in the full model.

Figures 4 and 5 show the point estimates and confidence intervals for our coefficients for the full and reduced model. As can be seen in Figure 4, the majority of the covariates do not seem to effect our predictions of the response substantially. In the reduced model we see that in both the elective and down syndrome case increasing religious importance and the presence of children was associated with less compassion, while older ages was associated with more compassion. None of the covariates were significant predictors for the rape or cancer situation.

One way in which we can visualize how well our model is working is demonstrated in Figures 7 and 8. These graphs shade in the regions where our θs (the cutoffs between our ordinal responses) were estimated to be , as well as plot the predicted latent variable score against the jittered observed response. For comparison, we fit a model to simulated data with strong covariates and extremely low noise to demonstrate what the ideal model would look like in Figure 6.

These images help us interpret our coefficients. Although the interpretation of the coefficients on the probability of each category is difficult, the coefficients can easily be interpreted on the latent variable scale. For example, in the full model we estimate the effect of religious importance on compassion for the elective case to be -0.298. We can interpret that as saying for a one standard deviation increase of religious importance (because we scaled the data), we expect the latent variable, $\tilde{Y}$, to decrease by -0.298. This doesn't lend itself to simple probabilistic statements because the widths of the intervals are uneven and the $\hat{\theta}$s are subject to sampling error. However, by looking at the plot we can see that the average interval is about one unit wide, so it would require about a three standard deviation increase in religious importance to decrease the ordinal response by one unit.

Several interesting results can be noted in the graphs. First, the lack of helpful predictors is clearly seen in the rape and cancer situations. The lack of spread indicates that none of the covariates are good at distinguishing the amount of compassion felt. The spread is much smaller in the reduced model because so much noise was being fitted by the unnecessary covariates. In these situations, the best guess for all the respondents is in the 'Quite a bit' range, whereas in the elective and down syndrome cases, the predictions will differ more dramatically depending on the covariates. These graphs also show the estimated width of the intervals. For example, we see that the 'moderate' category is smallest in the elective case, indicating

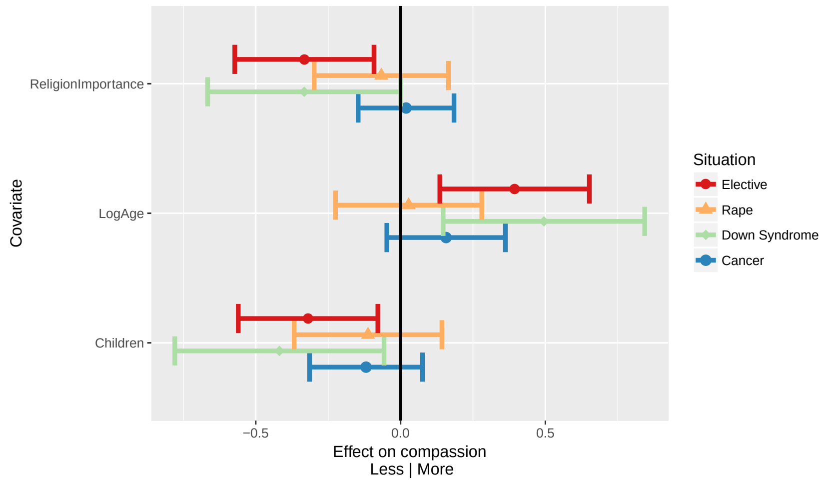

Figure 5: Coefficients for reduced model

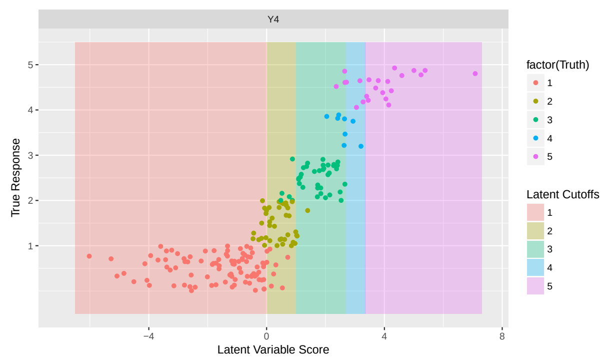

Figure 6: Plot generated with simulated data under ideal conditions

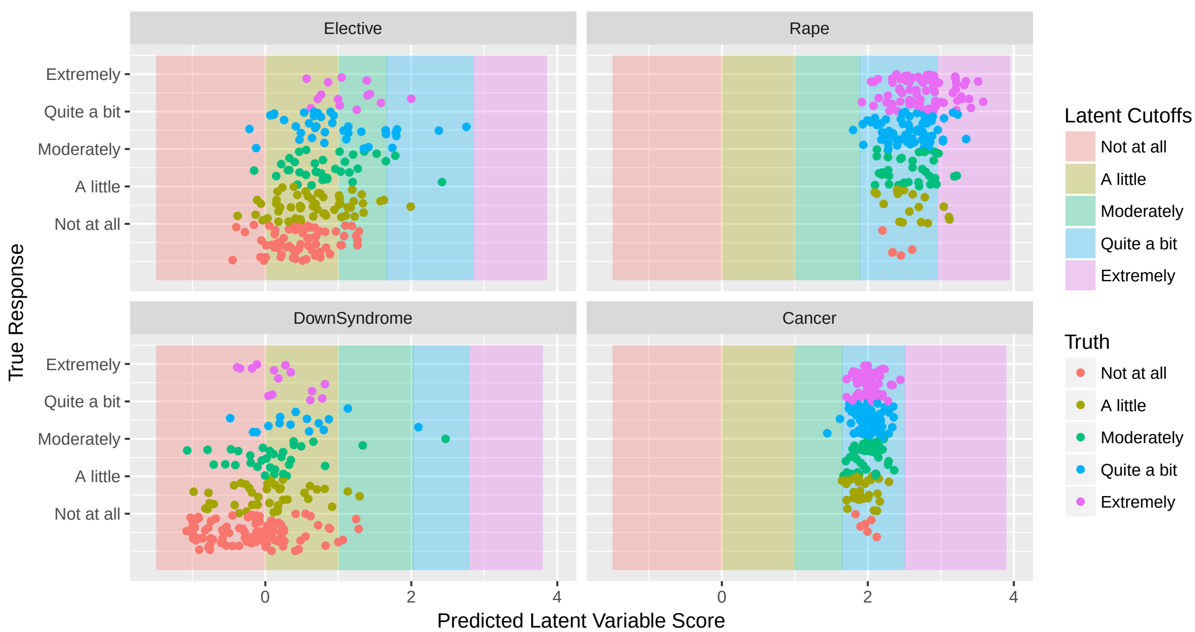

Figure 7: Estimated cutoffs and predicted scores for the latent variable for the full model

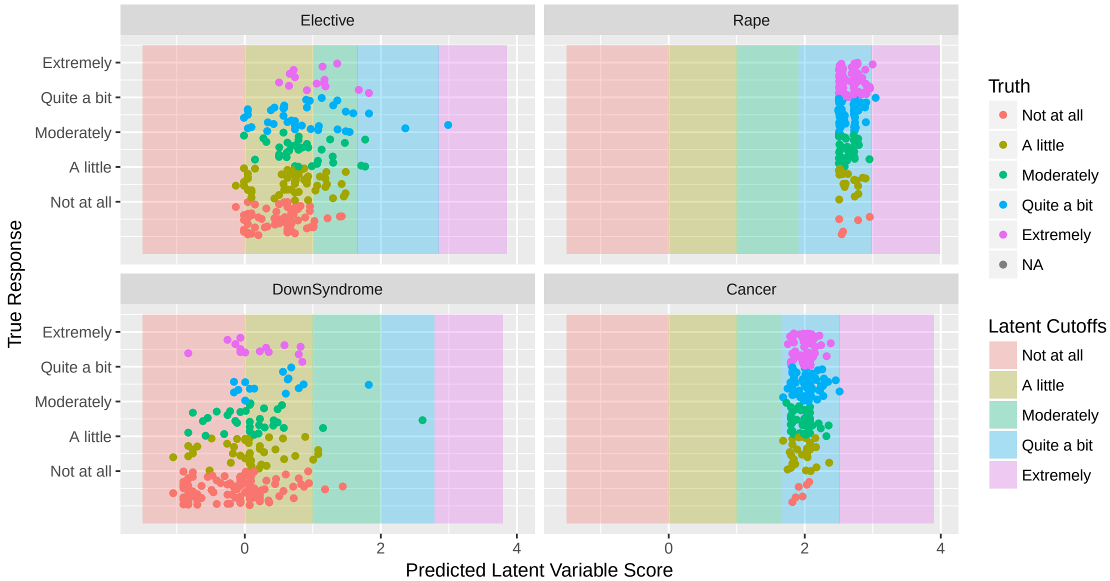

Figure 8: Estimated cutoffs and predicted scores for the latent variable for the reduced model

that it is rare for someone to feel a moderate amount of compassion for an elective abortion as compared to the other three scenarios.

One difficulty of this model is that one of its key assumptions is hard to verify. The model requires multivariate normally distributed errors on the latent variable scale, but we cannot observe the latent variables directly. We can approximate the latent variable errors by looking at the range of possible $\tilde{Y}$ for the observed Y and randomly selecting one to use when calculating $r_i = \tilde{y}_i - \hat{\tilde{y}}_i$. Figure 9 shows the histogram of these residuals. Clearly they are not ideal, but with a small sample size we do not expect to see perfect normality.

4.1 Error Analysis

One of the benefits of the multivariate analysis is the ability to analyze the correlation structure between the residual errors. Figures 10 and 11 show $\hat{\Sigma}$ converted to a correlation matrix for both the models and Table 1 gives the diagonal variances. Interestingly, the error correlation is estimated to be larger than the raw correlations in Figure 2. Also interesting is that all errors are positively correlated, indicating that

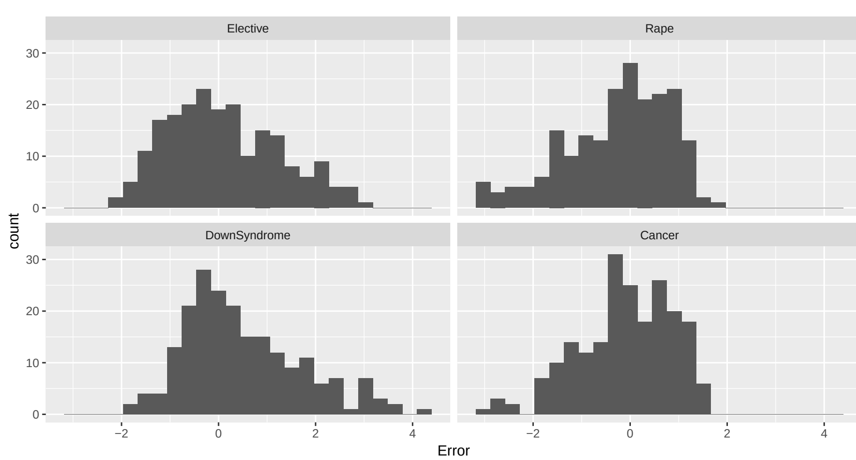

Figure 9: Approximate residuals for the latent variable

Model	Elective	Rape	Down Syndrome	Cancer
Full	1.71	1.69	3.13	1.05
Reduced	1.83	1.78	3.18	1.11

Table 1: Diagonal variances for $\hat{\Sigma}$

the information we are failing to capture generally boils down to a sense of compassion or lack thereof for abortion generally.

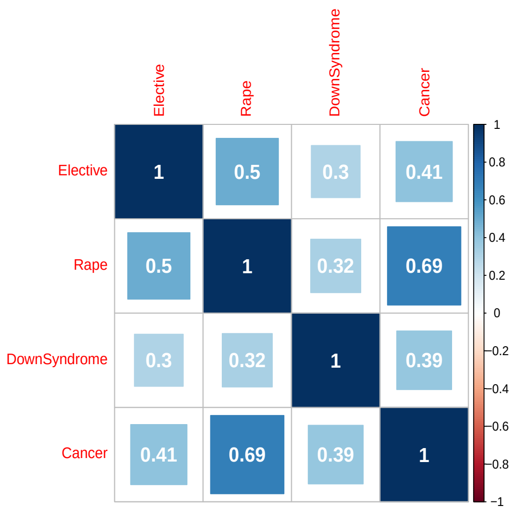

Figure 10: Correlation matrix for the full model errors

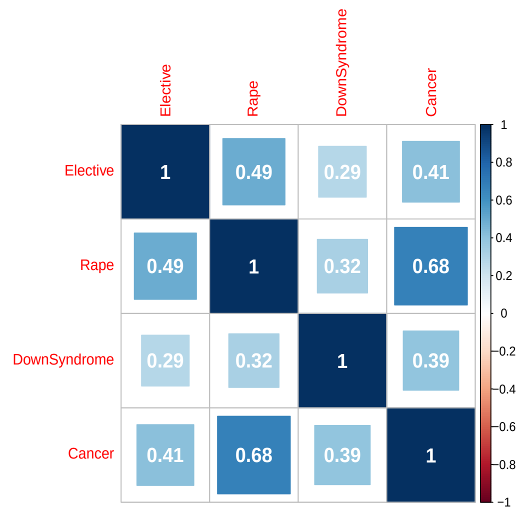

Figure 11: Reduced model errors

An interesting application of the error matrix is in data imputation. For example, two of our survey respondents did not respond to one situation apiece. Usually if we wanted to infer what they likely would have responded we could use the covariates to predict the response and that is it. However, in these two cases we also know what the error was between their expected response and their actual response in three of the four scenarios. The errors for each scenario are not independent, so we can adjust their predicted response by what we know about their other responses to get better predictions of the missing data.

To do this computation we will treat Σ, θ, and β as known. To get a true accounting that recognizes the uncertainty in the three sets of parameters either a bootstrapping or fully Bayesian approach would be appropriate, but we will demonstrate how the computation would proceed if the parameters were known.

Our first missing data point was in response to the elective abortion question. The respondent's answers to the rape, down syndrome, and cancer case were 'quite a bit', 'extremely', and 'moderately' respectively. We will use the midpoints of each of these intervals as their "true" latent variable score (although if we were bootstrapping it would be better to randomly select a point in each interval). Thus $\tilde{Y} = [NA, 2.4, 3.3, 1.3]$. By using this person's covariates we can calculate $\hat{\tilde{Y}} = [1.4, 2.8, 0.8, 2.1]$. Because we used the probit model with multivariate normal errors, we can use the following equations to calculate the conditional distribution of the first error given the values of the other three errors:

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{a} - \boldsymbol{\mu}_2\right) \tag{6}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

In our case, μ_1 and μ_2 are 0, and $a = \tilde{Y} - \hat{\tilde{Y}}$. Plugging in our numbers, we can calculate the updated error distribution for this person's elective abortion response. However, this computation is highly reliant on the assumption of multivariate normal errors, which Figure 9 identifies as questionable, so we won't interpret these too literally. The result of these formulas is:

$$\epsilon_1 \sim N(0.060, 1.246)$$

So in this case, while our predicted latent variable score only shifts up 0.06 from 1.375 to 1.435, our error variance drops from 1.712 in $\hat{\Sigma}$ to 1.246, so we become more confident in our prediction that this respondent will be moderately compassionate in the elective case.

In our second case the respondent failed to answer the question about down syndrome, but their responses to the elective, rape, and cancer scenarios were 'not at all', 'a little', and 'not at all' respectively. Following the steps above this means her latent variable errors were calculated as $[-0.82, -2.23, -2.33]$. In other words, she was far less compassionate in any situation then her covariates would have predicted. Thus, using the equations in 6, we can calculate the conditional error distribution for her down syndrome response:

$$\epsilon_3 \sim N(-1.47, 2.58)$$

Her covariates suggested a latent variable score of -0.83, so including this error we expect a mean latent variable score of -2.31 with a standard deviation of 1.61. Since the cutoff between 'not at all' compassionate and 'a little' compassionate is fixed to 0, we calculate the probability that her response is 'not at all' compassionate to be 92.5%, and the probability that she is more compassionate than 'a little' (which would require a latent score greater than 1) to be 2%. In contrast, the general population had 30% of respondents specify a response greater than 'a little'.

5 Conclusion

The multivariate ordinal model is a powerful tool to analyze survey data when the responses are Likert-scaled and closely linked. The framework provided in the mvord package is very flexible and easy to adapt to different situations. The probit model is conceptually easy and the normally distributed errors, while not necessarily always realistic, have nice theoretical properties.

In our analysis of the abortion data, we learned that simple covariates do not provide a great explanation of something as complex as the ethics of compassion and abortion. While variables such as age, religious identity, and having children do seem to effect your feelings about abortion, much of the variability in the data remains unexplained.

With regards to the four situations, some clear patterns emerged. Not only did the elective and down syndrome situations share similar effects for the covariates, but they also had highly correlated errors, suggesting that people view those two scenarios similarly. Likewise, the rape and cancer scenarios were similar in that the covariates did not seem at all predictive of people's responses, but people's responses in general were highly correlated, suggesting people might have the same ethical response to both those situations. In general, the strong positive correlations between all errors suggested that there is an unexplained trend toward compassion or lack of compassion in abortion situations.

From a social science research perspective, more investigation could be done into the characteristics that make rape and cancer similar as well as what characteristics make elective and down syndrome similar. Tightly controlling the wording of each scenario might allow a more fine-combed look at differences. If a more predictive model is wanted, gathering information on political party identification, religious group, region of the country, etc. would be beneficial.