Ordinal Regression Analysis: Factors Contributing to Job Satisfaction

By Angela Folz and Annelise Lynch

Motivation and Background

Interest

Survey data is heavily used in all industries because it is the most common way to collect data about customer and client satisfaction. For example, in the private sector, customer satisfaction leads to increase in conversions or decrease in churn, and ultimately more revenue.

More specifically, the General Social Survey (GSS) data is a valuable data source used by policy makers, economists, analysts, journalists, educators, students, healthcare administrators, and legislators. This social science data set was selected for this project to identify what factors contribute to job satisfaction, happiness, and financial satisfaction.

This project was motivated by the desire to focus on positive data around satisfaction and happiness, find factors that contribute to higher self-reported levels of satisfaction and happiness, and learn about ways these measurements are quantified. Methods of analyzing this data are less common than continuous regression. The knowledge and training required to analyze ordinal regression data will also be valuable skills to obtain.

Background data information

About the GSS survey data

The GSS project is within the National Opinion Research Center (NORC) at the University of Chicago. Most of the funding for the GSS project comes from the National Science Foundation. The purpose of the survey is to monitor societal change in the United States. The survey has been conducted from 1972-2018.

Below is an example of the data structure, and details of select variables used in this analysis.

Table 1: Example data

Respondent id number	Gss year for this respondent	Job or housework	Condition of health	Total family income	Respondents sex	Rs highest degree	Age of respondent	Number of hours usually work a week	How often does r find work stressful
1	2016	Mod. satisfied	Good	Refused	Male	Bachelor	47	Not applicable	No issp
2	2016	Very satisfied	Not applicable	$25000 or more	Male	High school	61	Not applicable	No issp
3	2016	Not applicable	Good	$25000 or more	Male	Bachelor	72	Not applicable	No issp

4	2016	Very satisfied	Good	Refused	Female	High school	43	Not applicable	No issp
1	2018	Mod. satisfied	Excellent	Refused	Female	Graduate	55	Not applicable	Hardly ever
2	2018	Very satisfied	Not applicable	$25000 or more	Female	Junior college	53	Not applicable	No issp
3	2018	Mod. satisfied	Poor	Refused	Male	High school	50	Not applicable	No issp

Table 2: Data details

Variable	**Survey question**	**Purpose in analysis**
Job or housework (satisfaction)	On the whole, how satisfied are you with the work you do--would you say you are very satisfied, moderately satisfied, a little dissatisfied, or very dissatisfied?	Response
Respondents Income	In which of these groups did your earnings from (OCCUPATION IN OCC) for last year--[the previous year]--fall? That is, before taxes or other deductions.	Predictor
Age of respondent	Respondents age	Predictor
Respondents sex	Respondents sex	Predictor
How often does respondent find work stressful	How often do you find your work stressful?	Predictor
Number of hours usually work a week	Number of hours usually work a week	Predictor
Condition of health	Would you say your own health, in general, is excellent, good, fair, or poor?	Predictor
Respondents highest degree	Respondents degree	Predictor

Data extraction

GSS has an extremely user friendly and accessible interface for building a data set and exporting it.

This was originally planned as a time series data analysis, and as such, variables were chosen that were available for every year recorded between 1972-2018 and were within the job satisfaction, general happiness, financial satisfaction, and demographic categories that GSS provided.

Ultimately, R was not able to handle the Excel data output, and a combination of .dct, .dat, and .txt files were used to import and transform the data before modeling it.

Limitations of survey collection and bias

The GSS website and documentation is very dense. This is not a complete list of limitations within the data but some large limitations that were discovered with regard to sampling that contribute to bias in the data:

- The survey is conducted in the United States only.
- Prior to 2006, the survey only sampled from the English speaking population. Starting in 2006, the survey was available to English and Spanish speakers. reference
- All surveys were conducted on-site at home “after 3:00 p.m. on weekdays or during the weekend or holidays”. reference
- Quota sampling was based on sex, age, and employment status only. reference

These limitations will lead to bias against people who work night shifts or off-hours, students, and US residents who do not speak English (or Spanish after 2006). While this is not something that is controlled for in this project, it is acknowledged as a known limitation of the data.

Motivation and questions about the data

At the onset of this project, the questions that were originally hoped to answer were:

- What factors contribute to job satisfaction?
- What factors contribute to overall happiness?
- What factors contribute to financial satisfaction?

And do these factors change over time?

After scouring through 76 predictor variables spanning 33 years, it became apparent that the project scope needed to be limited to just factors contributing to job satisfaction. Also, what appeared originally to be ordinal response time series data was determined to be just ordinal response data with no time series component. The assumptions that were made and later revealed as incorrect are detailed below in Methods.

Methods

Data cleansing and exploration

As expected, the data cleansing, parsing, and munging took the majority of the time and resources for this project. From direct experience, data is rarely ever provided in a format ready for analysis. The Jupyter notebook used to import, cleanse, and analyze the dataset is included as Appendix A. The full data cleansing can be found in this appendix, and is summarized below.

The data was provided in a double-encoded format: encoded column names and encoded values. The full list of encoded names and values can be found in **Appendix B** and **Appendix C**. Examples are provided below. Updating the code with the real values and real column names was very important for model development and interpretation, without them, the models

were impossible to interpret. This is exemplified in Table 4 below. When using the Code values, interpretation was near impossible since the buckets are not equally spaced, and start at 1 instead of 0.

Table 3: Column mappings

Coded Name	Label
id_	Respondent id number
year	Gss year for this respondent
race	Race of respondent
sex	Respondents sex
degree	Rs highest degree
educ	Highest year of school completed
age	Age of respondent
hrs2	Number of hours usually work a week
hrs1	Number of hours worked last week
rincome	Respondents income
income	Total family income

Table 4: Code value mappings

Variable name	Code	Label
rincome	99	No answer
rincome	98	Dont know
rincome	13	Refused
rincome	12	25000 or more
rincome	11	20000 to 24999
rincome	10	15000 to 19999
rincome	9	10000 to 14999
rincome	8	8000 to 9999
rincome	7	7000 to 7999
rincome	6	6000 to 6999
rincome	5	5000 to 5999
rincome	4	4000 to 4999
rincome	3	3000 to 3999
rincome	2	1000 to 2999
rincome	1	Lt 1000
rincome	0	Not applicable

The mapping values contained ASCII characters that R does not support. This needed to be cleansed manually. One example was one of the levels in the factor income: 1,000-2,000 needed to be changed to 1000to2000 for all of the models used in this analysis to not result in errors.

Some of the mapping values were factors, and some were continuous variables. Not all of these were intuitive. For example, income was in pre-assigned buckets (see Table 4 above), and it was expected that this would be a numerical input. The full list of factor versus continuous variables can be found in **Appendix B** in the “Variable Type” column.

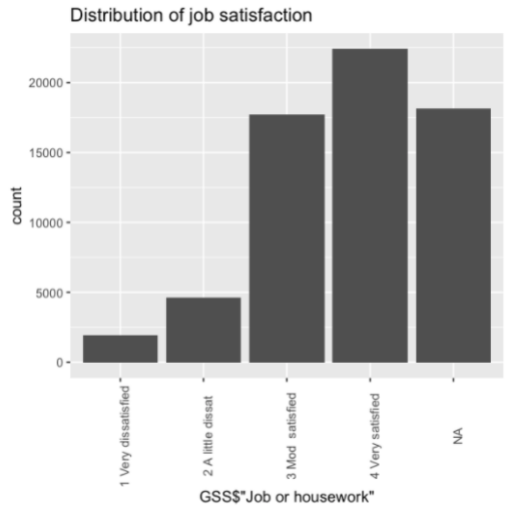
Many values needed to be encoded as NA, and this was different for each of the continuous and factor variables mentioned above. After many attempts and finding more values to encode as NA each time, the final resulting code looks like this:

Image 1: Applying NA to all columns for applicable responses

```
# turn into NANs: "Not applicable" and "No answer" for all columns
library(dplyr)
GSS = na_if(GSS, "Not applicable")
GSS = na_if(GSS, "No answer")
GSS = na_if(GSS, "Dont know")
GSS = na_if(GSS, "Refused")
GSS = na_if(GSS, "No issp")
GSS = na_if(GSS, "Cant choose")

# turn into NANs:
## 0 age, negative values in any of the continuous_factors
## 9, 99 or 98 in any of the continuous variables - these values map to 'No answer' and 'Dont know'
## Important: this cannot be done for Respondent ID because that makes null IDS, which is not valid

col_names = names(GSS)

for (i in col_names) {
    colname = toString(i)
    if(colname == 'Respondent id number'){
        next
    }
    GSS[colname] = na_if(GSS[colname], "9")
    GSS[colname] = na_if(GSS[colname], "99")
    GSS[colname] = na_if(GSS[colname], "98")
    GSS[colname] = na_if(GSS[colname], "13")
}

GSS$"Age of respondent" = na_if(GSS$"Age of respondent",0)

library(expss) # for using the lt() function below, means "less than"
for (i in continuous_fields) {
    colname = toString(i)
    GSS[colname] = na_if(GSS[colname], lt(0))
}

# Look at that cleaned data!
head(GSS)
```

After cleansing the data, the distributions of the response and predictor variables were explored. These can be seen in the figure below and in Appendix E Supporting Plots (note, data from all years was used in these plots).

Figure 1: Distribution of job satisfaction (NAs included for reference)



Assumptions

I) The data is linearly related.
See supporting plots in Appendix E and indicate linearity between predictors and response.

II) The response is categorical, so assumptions of normality and constant variance are not met, but are not required for the models used.

III) The response bias is accounted for in random sampling.
Since, unfortunately, respondent id is not persistent across years, this is not calculable to measure multiple responses from 1 person over time. This is also something that is well researched and one of the many response biases in survey results, some of which are listed here (resource):

- Demand bias
- Social desirability bias
- Dissent bias
- Acquiescence bias
- Extreme responses
- Neutral responding
- Personal bias
- Demand bias

All of these biases could exist in the dataset for this project. Some are unavoidable, such as demand bias, and we are trusting that the questions created by the GSS survey experts and the sampling used by GSS is robust enough to minimize the effects of these biases. However, it is possible that the noise created from these biases will lead to insignificant results.

Models (MVORD and POLR)

Multivariate ordinal regression (MVORD) model

Discoveries

I) The MVORD package is very computationally expensive.
A Macbook Pro was not able to process all GSS data, which is about 60MB in .dat file format. This is surprising, and thus the data was limited to just the most recent years when running the model: 2016 and 2018.

II) Respondent ID is not persistent across years
Respondent is not a true ID - it is unique only to the year and respondent. This was only discovered recently while proving that the assumptions for the MVORD package (detailed below) were met. For example, Respondent ID = 1 is just the first respondent for that year, not the same person year over year. This is poor experimental design as this is not intuitive. A true ID is unique to the object that it is labeling, which in this case, is an ID per respondent.

III) MVORD is not the correct package to use to model this data.
Unfortunately, this was only discovered after developing the models for the data. Originally, it was assumed that the Respondent ID was persistent (the same person) each year. This original understanding of the data led to learning about the MVORD package, which will be reviewed and discussed here. There is not much information online about this model as during model development googling errors + mvord yielded no results. It is exciting to learn about and present something that is not yet well established.

For the remainder of this section, please assume that the data is set up as was originally assumed: the Respondent ID is truly unique to 1 respondent and persists across years.

Understanding the MVORD model

The MVORD package is used to implement likelihood estimation for multivariate ordinal regression models with multivariate probit and logit links (source). This is perfect for our dataset because it takes a rating score, an id, and a time component to track responses for a specific id over time! Example 5 in this link illustrates exactly the model we wanted to implement and evaluate (where srhs is the response, id is a unique id, and time is year):

```
res_srhs <- mvord(formula = MMO(srhs, id, time) ~ 0 + factor(gender) +
    factor(race) + factor(education) + age,
    data = data_SRHS_long,
    threshold.constraints = rep(1, 8),
    coef.constraints = rep(1, 8),
    error.structure = cor_ar1(~ 1), link = mvlogit(),
    PL.lag = 2)
```

Which can be seen here using our data:

```
# using mvord library
# only using 2016 and 2018 data where none of the values listed below are NA
# mvord is not able to handle NAs

library(mvord)
mvord_3 <- mvord(data = GSS_job_sat,
                 formula = MMO(Job.or.housework,Respondent.id.number,Gss.year.for.this.respondent) ~
                     0 + Age.of.respondent,
    error.structure = cor_ar1(~ 1),
    link = mvlogit(),
    PL.lag = 2)

cat("made model")
print(summary(mvord_3))
```

Results

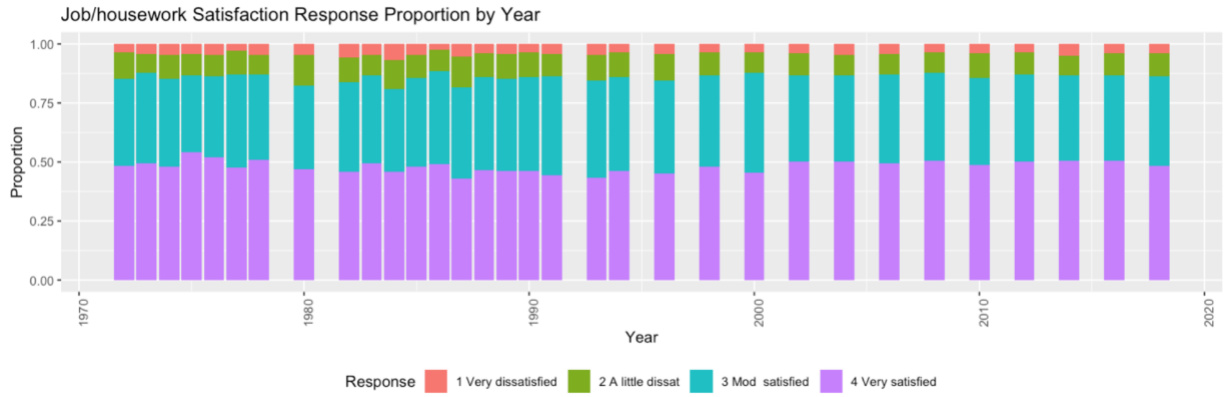
```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
Age.of.respondent 1 0.017938   0.030242  0.5932   0.5531
Age.of.respondent 2 0.018041   0.031369  0.5751   0.5652
```

The output of the logit model above (see full output in Appendix D), is interpreted as log-odds ratios. The results indicate that in 2018 being older increases the chances of higher satisfaction. And every year of increase in age the odds of job satisfaction increase by 1.82% (exp(0.018041) = 1.018204). Recall that this assumes the model is correct, which is not the case. (source)

Proportional odds logistic regression (POLR) model

Model-specific assumptions

I) Time series analysis is not necessary.
Since the data did not track respondent ids across different years, making MVORD impossible to use, we looked at the proportion of responses over time to see if there was a dramatic change over time. (We lacked the time and knowledge to approach this question through statistical methods, so we approached it graphically.) The figure below shows that the proportion of responses is fairly consistent throughout the entire time the data was collected. So for the OLS model a random year was picked (out of the years after the 2008 recession just in case that had an impact that we're not aware of) and methods were applied to data from that year (2010) only.



II) Proportional odds assumption (parallel regression assumption)
An assumption of ordinal logistic regression is that the coefficient for each pair of outcome groups is the same. Otherwise, more than one coefficient would be needed. (See the section on interpretation below for more information about the output of this model.) There are statistical methods for assessing whether this assumption is met, but arguments have been made against their reliability because they often reject the null hypothesis that the assumption is met when it should actually be accepted. There is a graphical method for checking this assumption, but we were unable to understand and implement it. Clearly, this is a necessary step before any conclusions can be made from the application of this model to our data! (source)

Understanding the POLR model

In this data, the response variable (job satisfaction) is categorical, and the categories are ordered:

1. Very dissatisfied
2. A little dissatisfied
3. Moderately satisfied
4. Very satisfied

So the appropriate model to use was ordinal logistic regression, specifically proportional odds logistic regression.

A proportional odds model models the probability of being in a particular category of the response or in any category below it versus being in a category above it. Mathematically, for a model with J categories of the response and p predictors,

$$logit[P(Y \leq j)] = \alpha_j - \sum_{i=1}^{p} \beta_i X_i$$

for j = 1, … , J-1 and i = 1, … , p, where j is a category of the response and i is a predictor variable. (source)

Things to note about this model:

- j = J is not included because the probability of being in the highest category is 1.
- The intercept term, α_j, changes with j. In other words, there will be a different intercept for each response category (see how this is interpreted below).
- Like binomial regression, the proportional odds model is in terms of the log odds of the probability, hence the logit function in the equation above. (source)

Implementing in R

In R, the polr() function in the MASS package implements this proportional odds logistic regression model. (source)

The output includes:

- Estimates, standard errors, and t values for each regression coefficient
- (No p-values are included, but these can be calculated using the t values and the standard normal distribution, assuming a large sample size)
- Estimates, standard errors, and t values for J-1 intercept terms
- Residual deviance and AIC

Interpretation

The data in this project includes both continuous and categorical predictors. The proportional odds model can take both into account.

Coefficients for continuous predictors are interpreted in the same way as for binomial regression. I.e., a coefficient of β_j for the jth (continuous) predictor can be interpreted to mean

that a one unit increase in the *j*th predictor is associated with an increase of β_j in the log odds of the response, adjusting for the other predictors and assuming the model is correct.

Coefficients for categorical predictors are a bit tricky to interpret. A coefficient of β_k corresponding to, for example, the gender predictor (where 0 indicates female and 1 indicates male) can be interpreted to mean that for respondents who are male, the odds of being more likely to be satisfied with their job is β_k times that of respondents who are female, adjusting for the other predictors and assuming the model is correct. See source for more details.

This model contains an intercept for each set of consecutive levels of the response. For our data with the job satisfaction response (assuming the model is correct), the intercept "Very dissatisfied | A little dissatisfied" can be interpreted as the log odds of being very dissatisfied versus being a little dissatisfied, moderately satisfied, or very satisfied (it corresponds to logit[P(Y ≤ 1)]). The intercept "A little dissatisfied | Moderately satisfied" can be interpreted as the log odds of being very dissatisfied or a little dissatisfied versus being moderately satisfied or very satisfied (corresponding to logit[P(Y ≤ 2)]). (source)

POLR hypotheses, methods, and results

Attempt 1
Hypothesis: Predictors directly related to the job (such as income and how often work is stressful) are the most significant predictors.

Method: Construct a POLR model using all variables (the full model) and a POLR model using only the following predictors, deemed to be the most directly related to the job:

- Does r supervise others at work in main job
- How often does r find work stressful
- R self-emp or works for somebody
- Labor force status
- Respondents income
- Number of employees: rs work site
- Number of hours worked last week

Note, the following predictors were excluded for the reasons listed:

- Rs industry code (naics 2007): too many categories
- Rs census occupation code (2010): too many categories
- Ever work as long as one year: errors likely because only 887 values are not NAs
- Number of hours usually work a week: errors, likely because only 40 values not NAs

Results: The full model failed to run. The problem didn't seem to be with certain variables because removing enough variables allowed it to run, regardless of which variables were removed.

Attempt 2

Hypothesis: Predictors directly related to the job (such as income and how often work is stressful) are more significant than predictors in other categories (such as marital status or place of residence when 16 years old).

Method: Predictors were split into the following categories:

- Job-related
- Family
- Personal history
- Other

Three "full" models were created by combining the job-related predictors with each of the other categories (e.g., one full model consisted of all of the job-related variables and all of the family variables). Four partial models were created, one for each category. The intention was to compare each full model with the job-related partial model using a likelihood ratio test.

However, first the residual deviance was used to calculate a p value, testing the hypothesis that the model is a good fit. This was done for each model (example below).

```
Residual Deviance:  1813.771
Degrees of freedom:  903
p value:  1.792271e-63
```

	Value	Std. Error	t value	p value
How.often.does.r.find.work.stressful2 Hardly ever	-0.62061815	0.340273509	-1.8238803	0.068
How.often.does.r.find.work.stressful3 Sometimes	-1.11250876	0.324453367	-3.4288711	0.001
How.often.does.r.find.work.stressful4 Often	-1.75394060	0.337253167	-5.2006646	0.000
How.often.does.r.find.work.stressful5 Always	-1.77274457	0.377413310	-4.6970908	0.000
R.self.emp.or.works.for.somebodySomeone else	-0.50779667	0.235556727	-2.1557298	0.031
Labor.force.statusWorking fulltime	0.30060746	0.228350023	1.3164328	0.188
Respondents.incomeb 1000 to 2999	-0.53542318	0.667820581	-0.8017470	0.423
Respondents.incomec 3000 to 3999	-0.75874750	0.674492576	-1.1249160	0.261
Respondents.incomed 4000 to 4999	-0.93175433	0.784212636	-1.1881399	0.235
Respondents.incomee 5000 to 5999	-0.30593150	0.733292118	-0.4172028	0.677

…

Results: All of the resulting p values were smaller than any commonly used significance level, indicating that none of the models were a good fit. Given this fact, it seemed irrelevant to compare two poorly-fitting models against each other.

Conclusions

While there were no significant relationships uncovered in this analysis, relationships may still exist. In lieu of significant results, this exploratory analysis led to learning about ordinal logistic regression and the MVORD and POLR models.

There are a few lessons learned from this project that are worthy of sharing. Primarily, while we enjoyed working with this data set, we should have followed an existing analysis using the GSS data instead of an exploratory analysis. This scope was too large for a semester project and it may have been good practice to try to replicate and reproduce another researcher's results. Also, in learning about ordinal logistic regression, we only found the proportional odds assumption later in the project. Proving these assumptions are met is complicated and in doing this project again, we would have started with proving that those assumptions were met with this dataset. (Perhaps a violation of this assumption is why there were no significant results from this model.) Finally, we did run into computationally limited issues with the mvord package. With the right data set, this model would require more hardware than a Macbook Pro to process a full data set.

If this research were to be extended, it should be turned into a true time series analysis (as it was originally intended). This could be done with outside data such as political events (elections), stock market indicators, large national events, and national disasters. This event data could be combined with satisfaction and happiness data to see how repeating events and trends affect satisfaction over time.

List of Appendices

- **Appendix A**: Jupyter notebook html
- **Appendix B**: Column name encoding with ASCII characters removed with continuous/factor types noted
- **Appendix C:** Value encodings with ASCII characters removed
- **Appendix D**: MVORD output
- **Appendix E**: Supporting plots and figures