“What is the concept of Ordinal Logistic Regression and how is it applied in SPSS for data analysis?”

Authored by

stats writer

June 29, 2024

RECOMMENDED CITATION

stats writer (2024). *“What is the concept of Ordinal Logistic Regression and how is it applied in SPSS for data analysis?”*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=157811

Ordinal Logistic Regression is a statistical method used to analyze and predict the relationship between a set of independent variables and an ordinal dependent variable. It is commonly used in data analysis to determine the likelihood of an outcome falling into a certain category, such as low, medium, or high. In SPSS, this concept is applied by fitting a model to the data and estimating the coefficients for each independent variable, which can then be used to make predictions about the outcome. This method allows for the analysis of data with multiple categories and can provide valuable insights for decision making.

Ordinal Logistic Regression | SPSS Data Analysis Examples

Version info: Code for this page was tested in IBM SPSS 20.

Please note: The purpose of this page is to show how to use various data analysis commands. It does not cover all aspects of the research process which researchers are expected to do. In particular, it does not cover data cleaning and checking, verification of assumptions, model diagnostics and potential follow-up analyses.

Examples of ordered logistic regression

Example 1: A marketing research firm wants to investigate what factors influence the size of soda

(small, medium, large or extra large) that people order at a fast-food chain. These factors may include what type of sandwich is ordered (burger or chicken), whether or not fries are also ordered, and age of the consumer. While the outcome variable, size of soda, is obviously ordered, the difference between the various sizes is not consistent. The difference between small and medium is 10 ounces, between medium and large 8, and between large and extra large 12.

Example 2: A researcher is interested in what factors influence medaling in Olympic swimming. Relevant predictors include at training hours, diet, age, and popularity of swimming in the athlete's home country. The researcher believes that the distance between gold and silver is larger than the distance between silver and bronze.

Example 3: A study looks at factors that influence the decision of whether to apply to graduate school. College juniors are asked if they are unlikely, somewhat likely, or very likely to apply to graduate school. Hence, our outcome variable has three categories. Data on parental educational status, whether the undergraduate institution is public or private, and current GPA is also collected. The researchers have reason to believe that the "distances" between these three points are not equal. For example, the "distance" between "unlikely" and "somewhat likely" may be shorter than the distance between "somewhat likely" and "very likely".

Description of the data

For our data analysis below, we are going to expand on Example 3 about applying to graduate school. We have simulated some data for this example and it can be obtained from here: ologit.sav

This hypothetical data set has a three-level variable called apply (coded 0, 1, 2), that we will use as our outcome variable. We also have three variables that we will use as predictors: pared, which is a 0/1 variable indicating whether at least one parent has a graduate degree; public, which is a 0/1 variable where 1 indicates that the undergraduate institution is public and 0 private, and gpa, which is the student's grade point average.

Let's start with the descriptive statistics of these variables.

get file "D:dataologit.sav".
freq var = apply pared public.

apply

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	220	55.0	55.0	55.0
	1.00	140	35.0	35.0	90.0
	2.00	40	10.0	10.0	100.0
	Total	400	100.0	100.0	

pared

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	337	84.3	84.3	84.3
	1.00	63	15.8	15.8	100.0
	Total	400	100.0	100.0	

public

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	343	85.8	85.8	85.8
	1.00	57	14.2	14.2	100.0
	Total	400	100.0	100.0	

descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
pared	400	.00	1.00	.1575	.36473
public	400	.00	1.00	.1425	.35000
gpa	400	1.90	4.00	2.9989	.39794
Valid N (listwise)	400				

var = gpa.

Analysis methods you might consider

Below is a list of some analysis methods you may have encountered.

Some of the methods listed are quite reasonable while

others have either
fallen out of favor or have limitations.

Ordered logistic regression

Before we run our ordinal logistic model, we will see if any cells are empty
or extremely small. If any are, we may have difficulty running our model.
There are two ways in SPSS that we can do this. The first way is to make
simple crosstabs. The second way is to use the cellinfo option on
the /print subcommand. You should use the cellinfo option
only with categorical predictor variables; the table will be long and difficult
to interpret if you include continuous predictors.

crosstabs
/tables = apply by pared
/tables = apply by public.

apply * pared Crosstabulation

Count

		pared		
		.00	1.00	Total
apply	.00	200	20	220
	1.00	110	30	140
	2.00	27	13	40
Total		337	63	400

apply * public Crosstabulation

Count

		public		
		.00	1.00	Total
apply	.00	189	31	220
	1.00	124	16	140
	2.00	30	10	40
Total		343	57	400

plum apply with pared public

/link = logit

Cell Information

Frequency

pared	public		apply .00	apply 1.00	apply 2.00
.00	.00	Observed	175	98	20
		Expected	174.640	95.387	22.973
		Pearson Residual	.043	.326	-.646
	1.00	Observed	25	12	7
		Expected	25.168	15.054	3.778
		Pearson Residual	-.051	-.970	1.734
1.00	.00	Observed	14	26	10
		Expected	16.233	23.414	10.352
		Pearson Residual	-.674	.733	-.123
	1.00	Observed	6	4	3
		Expected	3.944	6.147	2.909
		Pearson Residual	1.241	-1.193	.061

Link function: Logit.

/print = cellinfo.

None of the cells is too small or empty (has no cases), so we will run our model. In the syntax below, we have included the link = logit subcommand, even though it is the default, just to remind ourselves that we are using the logit link function. Also note that if you do not include the print subcommand, only the Case Processing Summary table is provided in the output.

plum apply with pared public gpa
/link = logit
/print = parameter summary.

Case Processing Summary

		N	Marginal Percentage
apply	.00	220	55.0%
	1.00	140	35.0%
	2.00	40	10.0%
Valid		400	100.0%
Missing		0	
Total		400	

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	557.272			
Final	533.091	24.180	3	.000

Link function: Logit.

Pseudo R-Square

Cox and Snell	.059
Nagelkerke	.070
McFadden	.033

Link function: Logit.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[apply = .00]	2.203	.784	7.890	1	.005	.666	3.741
	[apply = 1.00]	4.299	.809	28.224	1	.000	2.713	5.885
Location	pared	1.048	.268	15.231	1	.000	.522	1.574
	public	-.059	.289	.041	1	.839	-.624	.507
	gpa	.616	.263	5.499	1	.019	.101	1.130

Link function: Logit.

In the Case Processing Summary table, we see the number and percentage of cases in each level of our response variable. These numbers look fine, but we would be concerned if

one level
had very few cases in it. We also see that all 400 observations in our
data set were used in the analysis. Fewer observations would have been
used if any of our variables had missing values. By default, SPSS does a listwise
deletion of cases with missing values. Next we see the Model Fitting
Information table, which gives the -2 log likelihood for the intercept-only and
final models. The -2 log likelihood can be used in comparisons of nested
models, but we won't show an example of that here.

In the Parameter Estimates table we see the coefficients, their standard errors, the
Wald test and associated p-values (Sig.), and the 95% confidence interval of the coefficients.
Both pared and gpa are statistically significant; public is
not.& So for pared, we would say that for a one unit
increase in pared (i.e., going from 0 to 1), we expect a 1.05 increase in

the ordered log odds of being in a higher level of apply, given all of the other variables in the model are held constant. For gpa, we would say that for a one unit increase in gpa, we would expect a 0.62 increase in the log odds of being in a higher level of apply, given that all of the other variables in the model are held constant. The thresholds are shown at the top of the parameter estimates output, and they indicate where the latent variable is cut to make the three groups that we observe in our data. Note that this latent variable is continuous.

In general, these are not used in the interpretation of the results. Some statistical packages call the thresholds "cutpoints" (thresholds and cutpoints are the same thing); other packages, such as SAS report intercepts, which are the negative of the thresholds. In this example, the intercepts would be -2.203 and -4.299. For further information, please see the Stata FAQ:

How can I convert Stata's parameterization of ordered probit and logistic models to one in which a constant is estimated?

As of version 15 of SPSS, you cannot directly obtain the proportional odds ratios from SPSS. You can either use the SPSS Output Management System (OMS) to capture the parameter estimates and exponentiate them, or you can calculate them by hand. Please see Ordinal Regression by Marija J. Norusis for examples of how to do this. The commands for using OMS and calculating the proportional odds ratios is shown below. For more information on how to use OMS, please see our SPSS FAQ: How can I output my results to a data file in SPSS? Please note that the single quotes in the square brackets are important, and you will get an error message if they are omitted or unbalanced.

oms select tables

```
/destination format = sav outfile = "D:ologit_results.sav"
/if commands = subtypes = .

plum apply with pared public gpa
/link = logit
/print = parameter.

omsend.

get file "D:ologit_results.sav".
rename variables Var2 = Predictor_Variables.
* the next command deletes the thresholds from the data set.
select if Var1 = "Location".
exe.
* the command below removes unnessary variables from the data set.
* transformations cannot be pending for the command below to work, so
* the exe.
* above is necessary.
delete variables Command_ Subtype_ Label_ Var1.
compute expb = exp(Estimate).
compute Lower_95_CI = exp(LowerBound).
compute Upper_95_CI = exp(UpperBound).
```

execute.

	Predictor_Variables	Estimate	Std.Error	Wald	df	Sig	LowerBound	UpperBound	expb	Lower_95_CI	Upper_95_CI
1	pared	1.048	.268	15.231	1	.000	.522	1.574	2.85	1.68	4.83
2	public	-.059	.289	.041	1	.839	-.624	.507	.94	.54	1.66
3	gpa	.616	.263	5.499	1	.019	.101	1.130	1.85	1.11	3.10
4											

In the column expb we see the results presented as proportional odds ratios (the coefficient exponentiated). We have also calculated the lower and upper 95% confidence interval. We would interpret these pretty much as we would odds ratios from a binary logistic regression. For pared, we would say that for a one unit increase in pared, i.e., going from 0 to 1, the odds of high apply versus the combined middle and low categories are 2.85 greater, given that all of the other variables in the model are held constant. Likewise, the odds of the combined middle and high categories versus low apply is 2.85 times greater, given that all of the other variables in the model are

held constant. For a one unit increase in gpa, the odds of the low and middle categories of apply versus the high category of apply are 1.85 times greater, given that the other variables in the model are held constant. Because of the proportional odds assumption (see below for more explanation), the same increase, 1.85 times, is found between low apply and the combined categories of middle and high apply.

One of the assumptions underlying ordered logistic (and ordered probit) regression is that the relationship between each pair of outcome groups is the same. In other words, ordered logistic regression assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and

all higher categories,
etc. This is called the proportional odds assumption or the parallel
regression assumption. Because the relationship between all pairs of
groups is the same, there is only one set of coefficients (only one model).
If this was not the case, we would need different models to describe the
relationship between each pair of outcome groups. We need to test the
proportional odds assumption, and we can use the tparallel option on the
print subcommand. The null hypothesis of this chi-square test is
that there is no difference in the coefficients between models, so we hope to
get a non-significant result.


```
plum apply with pared public gpa
/link = logit
/print = tparallel.
```

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	533.091			
General	529.077	4.014	3	.260

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

The above test indicates that we have not violated the proportional odds assumption. If the proportional odds assumption was violated, we may want to go with multinomial logistic regression.

We we use these formulae to calculate the predicted probabilities for each level of the outcome, apply. Predicted probabilities are usually easier to understand than the coefficients or the odds ratios.

$$P(Y = 2) = left(frac{1}{1 + e^{-(a_{2}+b_{1}x_{1} + b_{2}x_{2} + b_{3}x_{3})}}right)$$
$$P(Y = 1) = left(frac{1}{1 + e^{-(a_{1}+b_{1}x_{1} + b_{2}x_{2} + b_{3}x_{3})}}right) - P(Y = 2)$$
$$P(Y = 0) = 1 - P(Y = 1) - P(Y = 2)$$

We will calculate the predicted probabilities using SPSS' Matrix language.
We will use pared as an example with a categorical predictor. Here we will
see how the probabilities of membership to each category of apply change
as we vary pared and hold the other variable at their means.
As you can see, the predicted probability of being in the lowest category of apply
is 0.59 if neither parent has a graduate
level education and 0.34 otherwise. For the middle category of apply, the
predicted probabilities are 0.33 and 0.47, and for the highest category of
apply, 0.078 and 0.196 (annotations were added to the output for clarity). Hence, if neither of a respondent's parents
have a graduate level education, the predicted probability of applying to
graduate school decreases. Note that the intercepts are the negatives of the
thresholds. For a more detailed explanation of how to interpret the predicted probabilities and its relation to

the odds ratio, please refer to FAQ: How do I interpret the coefficients in an ordinal logistic regression?

```
Matrix.
* intercept1 intercept2 pared public gpa.
* these coefficients are taken from the output.
compute b = {-2.203 ; -4.299 ; 1.048 ; -.059 ; .616}.
* overall design matrix including means of public and gpa.
compute x = {{0, 1, 0; 0, 1, 1}, make(2, 1, .1425), make(2, 1, 2.998925)}.
compute p3 = 1/(1 + exp(-x * b)).
* overall design matrix including means of public and gpa.
compute x = {{1, 0, 0; 1, 0, 1}, make(2, 1, .1425), make(2, 1, 2.998925)}.
compute p2 = (1/(1 + exp(-x * b))) - p3.
compute p1 = make(NROW(p2), 1, 1) - p2 - p3.
compute p = {p1, p2, p3}.
print p / FORMAT = F5.4 / title = "Predicted Probabilities for Outcomes 0 1 2 for pared 0 1 at means".
End Matrix.
```

Run MATRIX procedure:

Predicted Probabilities for Outcomes 0 1 2 for pared 0 1 at means
(apply=0) (apply=1) (apply=2)
(pared=0) .5900 .3313 .0787
(pared=1) .3354 .4687 .1959

------ END MATRIX -----

Below, we see the predicted probabilities for gpa at 2, 3 and 4. You can see that the
predicted probability increases for both the middle and highest categories of
apply as gpa increases (annotations were added to the output for clarity). For a more detailed explanation of how to interpret the predicted probabilities and its relation to the odds ratio, please refer to FAQ: How do I interpret the coefficients in an ordinal logistic regression?

Matrix.
*** intercept1 intercept2 pared public gpa.**
*** these coefficients are taken from the output.**
compute b = {-2.203 ; -4.299 ; 1.048 ; -.059 ; .616}.
*** overall design matrix including means of pared and**

```
public.
compute x = {make(3, 1, 0), make(3, 1, 1), make(3, 1, .1575), make(3, 1, .1425), {2; 3; 4}}.
compute p3 = 1/(1 + exp(-x * b)).
* overall design matrix including means of pared and public.
compute x = {make(3, 1, 1), make(3, 1, 0), make(3, 1, .1575), make(3, 1, .1425), {2; 3; 4}}.
compute p2 = (1/(1 + exp(-x * b))) - p3.
compute p1 = make(NROW(p2), 1, 1) - p2 - p3.
compute p = {p1, p2, p3}.
print p / FORMAT = F5.4 / title = "Predicted Probabilities for Outcomes 0 1 2 for gpa 2 3 4 at means".
End Matrix.
```

Run MATRIX procedure:

```
Predicted Probabilities for Outcomes 0 1 2 for gpa 2 3 4 at means
(apply=0) (apply=1) (apply=2)
(gpa=2) .6930 .2553 .0516
(gpa=3) .5494 .3590 .0916
(gpa=4) .3971 .4456 .1573

------ END MATRIX -----
```

Things to consider

References