Homework 6

ST623

Nick Sun

November 17, 2019

Question 1

Introduction

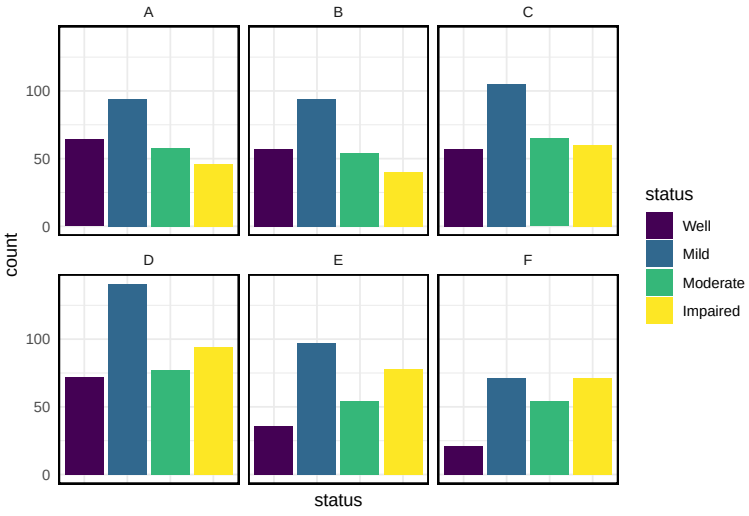
For this problem, we will be analyzing data on the mental health status children as a response to the socioeconomic status (SES) of the parents. The SES groups are ordered A-F with A being the most well off and the mental health status groups are ordered from "Well" to "Impaired".

The first few rows of the data are displayed below:

SES	status	count
A	Well	64
A	Mild	94
A	Moderate	58
A	Impaired	46
B	Well	57
B	Mild	94

There is a relatively equal sized sample for each of the SES groups, although there are definitely more families in group "D" than the rest.

Let's make a quick visual of our data.



There lower SES grous appear to have lower relative proportions of children who are classified as "Well". The higher SES groups appear to have lower relative proportions of "Impaired" children.

Model

Since the response is an ordered categorical variable, we will use an ordinal regression model. This can be done with the `polr` function from the **MASS** package.

We begin by using a logistic link function. The summary of this model is provided below:

	Value	Std..Error	t.value	p.val
SESB	-0.01697	0.1608	-0.1056	0.9159
SESC	0.2082	0.1548	1.345	0.1787
SESD	0.299	0.1458	2.051	0.0403
SESE	0.5668	0.1584	3.578	0.0003459
SESF	0.8239	0.1662	4.957	7.173e-07
Well|Mild	-1.204	0.1193	-10.09	6.015e-24
Mild|Moderate	0.4953	0.115	4.307	1.658e-05
Moderate|Impaired	1.504	0.1203	12.51	6.885e-36

Checking the performance of the `probit` and `cloglog` link functions shows that there is only a very slight difference in AIC between the three link functions. Therefore, in the interest of more familiar interpretation, we will stick with logistic regression since it is more familiar to most people.

logistic	probit	cloglog
4449	4447	4450

The coefficients of each SES group in our ordinal logistic model grow from "A" through "F". In creating our data, we ordered the response variable from "Well" to "Imparied", so the increasing coefficients indicate that the odds of a student being in the lower health categories increases as SES decreases.

Let's check the fitted values of this model.

	Well	Mild	Moderate	Impaired
A	0.2308	0.3906	0.1968	0.1818
B	0.2338	0.3915	0.1954	0.1793
C	0.1959	0.3754	0.2139	0.2149
D	0.182	0.3669	0.2205	0.2306
E	0.1455	0.3367	0.2364	0.2814
F	0.1163	0.3023	0.2452	0.3362

The rows are the fitted proportions within each group, so each row sums to 1. Indeed, the "Well" group tends to have a lower relative proportion in the lower SES groups than the higher ones. The "Impaired" group by contrast tends to have a higher proportion in the lower SES groups.

Collapsing Categories

We can see if the performance is improved by collapsing response categories together, for example, what if we combined "Well" and "Mild" or "Moderate" and "Impaired". After collapsing pairs of consectutive categories together and testing if we get an appreciably better fit, collapsing "Mild" and "Moderate" into a single "Mild" response category gives the lowest AIC value.

Here is the summary of that model.

Re-fitting to get Hessian

Call: polr(formula = collapsed_status ~ SES, data = collapsed_data, weights = collapsed_count, method = "logistic")

Table 5: Coeficients

	Value	Std. Error	t value
SESB	0.001079	0.1743	0.006188
SESC	0.2449	0.1683	1.455
SESD	0.3916	0.1585	2.47
SESE	0.7006	0.172	4.072
SESF	0.9091	0.1803	5.043

Table 6: Intercepts

	Value	Std. Error	t value
Well|Mild	-1.15	0.1257	-9.148
Mild|Impaired	1.575	0.1295	12.16

Residual Deviance: 3170.064

AIC: 3184.064

One thing that is immediately interesting is that the AIC value of the collapsed model drops to 3184.064, significantly lower than the model with 4 response categories. Doing a formal likelihood ratio test using the `lmtest` package (note that `anova.polr` will not work here since the data is not of the same dimensions) produces a miniscule p-value ($< 2.2e\text{-}16$). This low p-value tells us that the collapsed status ordinal model performs significantly better than the ordinal model we used previously.

Looking at the fitted values, this model again demonstrates a pronounced difference in the relative proportions between the higher SES groups "A" and "B" and the lower SES groups.

	Well	Mild	Impaired
A	0.2405	0.588	0.1715
B	0.2403	0.588	0.1716
C	0.1987	0.5922	0.2091
D	0.1763	0.5892	0.2344
E	0.1358	0.5699	0.2943
F	0.1132	0.5475	0.3394

Conclusion

- In general, mental health in children declines as SES declines.
  - Using our full model with 4 response categories, we can interpret our coefficients to get specific odds ratios.
  - There is no statistically significant difference between categories "A" and "B", but the categories "C" through "E" are significantly different from A.
- The best performing model collapses the "Mild" and "Moderate" response categories together.
  - We should be cautious about interpreting these coefficients – the literature is unclear about whether we can perform valid inference after collapsing response categories together.