Introduction

Dementia is characterized as the loss of cognitive function, and is strongly associated with the elderly in popular understandings[Ric06]. It frequently interferes with a person's ability to reason, remember and function on a daily basis. As the baby boom generation ages, and as a reduction in sudden mortality, or preventable causes younger in life changes the demographics of the world population, understanding the social and medical factors that play a role in the severity of dementia will be important.

The Data

Our data were compiled by Nicky Wakim, from the National Alzheimer's Consulting Center. Variables included cognitive status (CDR) which is our main variable of interest, race, which was rebalanced due to undersampling in the original sample. Our research questions is "Is education level associated with cognitive impairment due to dementia later in life?" It seems that there is something inherent to dementia in the way that education changes our brains [Car10]. We are also interested in the interaction between education and residence type inform cognitive impairment; perhaps there is some type of interaction between education and socioeconomic status and the social environment of the elderly. Finally, dementia also seems to be something that people develop as they age, so we wanted to consider an interaction between age and controlling for age, as well as education level, as with the advent of the GI Bill, many white Americans started getting more and more education who are now at the age where they tend to develop dementia.

Initial Model

Since there were limited variables in the dataset, we began by examining the literature for all included variables: marital status, gender, family status (regarding dementia diagnoses), age of diagnosis, race, education, residential status, clinical dementia rating, and depression. After reading several papers about the ways that certain variables could affect the relationship between outcome (clinical dementia rating) and exposure (education level), we created a Directed Acyclic Graph to ensure that we were capturing all the different ways that confounding could happen in our data set.

In Figure 1, the relationship of interest is Education on Clinical Dementia Rating (relationships isolated via [Cle14; Már10] for Family Status, [Dor12; Sar14] for Evaluation Age, [C C18; Gla95; Mar06] for Race, [Emi11; G A98; D W91] for Education, [Ada14; Pie09; She14] for Residential Status, [Gen15] for gender and [Cat09] for marriage). The variables outlined in red are the variables that we decided to control for because they are confounded with the relationship of interest. There are some variables that we wish we could have observed. For example, we know that there is a relationship between Race, education, income, and residential status. However, the data does not have an income or net worth variable. Access to dementia screening tests is also not recorded but that would also be interesting information to have. Note that the response variable is either 0, 0.5, or 1. This means that we will fit a multinomial regression model to the data.

Methods

While model fitting, we had to do some improvisation with variables and data cleaning. For example, there were other public residential facilities and "other" values in residence that were combined into "non-private" residence. Also, marriage status were coded into binary classes indicating married or not, instead of imbalanced groups describing detailed marriage status. Coding the previous two variables into binary outcome does not affect the significant association (Chi-square test) between the two variables and the CDR level, which is evidence to support this binary encoding. Education was also treated as a continuous variable in the original dataset, but we decided to reclassify it and treat it as a categorical dataset due to an imbalance in proportions. See Table 3 for descriptive statistics of variables after preprocessing. We first fit a proportional odds model as seen below based on the literature review, and our question of interest. Our final model (M1) is:

$$
\begin{aligned}
[M1]\ \ g(y) = \beta_{0|.5} + \beta_{.5|1} + \beta_1\ \text{Educ}_B + \beta_2\ \text{Educ}_C + \beta_3\ \text{Educ}_D + \beta_4\ \text{Male} + \beta_5\ \text{Married} + \beta_6\ \text{Depressed} + \beta_7\ \text{White} \\
+ \beta_8\ \text{EvalAge} + \beta_9\ \text{EvalAge} + \beta_{10}\ \text{Educ}_B \times\ \text{EvalAge} + \beta_{11}\ \text{Educ}_C \times\ \text{EvalAge} + \beta_{12}\ \text{Educ}_D \times \text{Private} \\
+ \beta_{13}\ \text{Educ}_B \times\ \text{Private} + \beta_{14}\ \text{Educ}_C \times\ \text{Private} + \beta_{15}\ \text{Educ}_D \times \text{Private}
\end{aligned}
$$

where g is the cumulative logit link function, and $y \in \{0, .5, 1\}$ which are the clinical dementia ratings. See Table 1 and Figure 5 in the appendix for model output. There is no significant multicollinearity among covariates from the heat plot as shown in Figure 4. The result from the heat-plot also agrees with the result of gvif as shown in Figure 6.

However, when we ran a likelihood ratio test between the proportional odds model and a cumulative logit model, the test (p-value $< 2.2 \times 10^{-16}$) indicated that the proportionality assumption was violated. When we attempted to then run a cumulative logit odds model `R` immediately threw a numerical error (see Appendix, Figure 8 for more detail). Additionally, we checked the log likelihood for the model, which is -2615.013. The log likelihood is much smaller than the log likelihood of model with `parallel=True`, whose loglikehood is -707.422. This is most likely happening because there are far too many slopes of lines for our categorical data to be interpreted with the `vglm` package. This is a common error with vectorized generalized linear models.

We decided to evaluate each variable to decide if there was one in particular that was causing a numerical error. We added variables to our man effects model one at a time, and found that age at evaluation and marital status were causing issues with numerical evaluation. It is unsurprising that age of evaluation is causing issues; there are many levels to this variable that may be causing the estimation procedure to break down. However, marriage only has two levels and is relatively balanced, so it came as a surprise that such a variable caused problems.

Ultimately, the model closest to the model indicated by Figure 1 that we could fit without estimation errors is:

$$[M2] \quad g(y) = \beta_{0|.5} + \beta_{.5|1} + \beta_1 \text{ Educ}_{B,1} + \beta_2 \text{ Educ}_{C,1} + \beta_3 \text{ Educ}_{D,1} + \beta_4 \text{ Educ}_{B,2} + \beta_5 \text{ Educ}_{C,2} + \beta_6 \text{ Educ}_{D,2} + \beta_7 \text{ Male}_1 + \beta_8 \text{ Male}_2 + \beta_9 \text{ Depressed}_1 + \beta_{10} \text{ Depressed}_2 + \beta_{11} \text{ White}_1 + \beta_{12} \text{ White}_2$$

Again, g is the cumulative logit link function, and y is clinical dementia ratings.

Diagnostics

As we discussed in class, there are very few diagnostic test for multinomial models. However, we did run the Akaike's information criterion (AIC) and found the AIC of the proportional odds model to be 1433.598 on 16 parameters. Additionally, a Likelihood Ratio Test of final proportional odds model compared to the null model indicated that we prefer the full model (p-value of $< 2.2 \times 10^{-16}$, the smallest value that `R` can hold). The McFadden's pseudo-R^2 is 0.262. Finally, the Hosmer-Lemeschow test indicated that our model (M1) fits well, but one of the weakness of the test is that we have to specify a g-value which can be fairly arbitrary. Our model falls prey to that weakness, with small changes in g yielding large changes in p-value (See Table 2). However, in our case, the results of HL-test are consistent regardless of the g-value. This provides more evidence that our model (M1) is fairly appropriate for the data.

Results

As we noted above, there were problems with the assumptions we made when fitting the proportional odds model (M1), but fitting a nominal model for the same covariates meant that we ran into numerical issues. This resulted in a choice that we had to make: choose a model that does not violate assumptions, or choose a model that we know does not account for all confounding (M2). Figure 2 shows the new relationships that we are unable to totally control for when we fail to include Age at Evaluation and Marital Status in our model. The paths in blue, green and purple represent new ways for Education to have an affect on the outcome that we are unable to control for, thus failing to isolate our outcome of interest.

Because this failure to isolate the association between our exposure and outcome seems to be a large violation and thus invalidating all of our results, we decided to choose the proportional odds model (M1) that may violate the proportionality assumption, but at least we are able to control for confounders in this model. The estimates that we get out might be biased due to a poor fit of model, but as of now, it is the better option than the model with parallel is False.

As it stands, the model that we decided to work with (M1) indicates that the third quartile of education (complete college) leads to a better outcome than the other three quartiles of education. Specifically, the odds ratio of having no or mild cognitive impairment comparing the third to first quartiles of education is 20.90 (p-value $< .05$). This is incredibly significant and indicates that there is an association between higher education and dementia for those who complete college and/or a masters compared to those that don't attend college. This indicates that an association likely exists between higher

education and dementia. Figure 7 depicts the difference in probability of being a member of no to moderate cognitive impairment and high impairment.

Given that dementia is largely associated with the elderly, we also wanted to gain some control for age, as American education patterns have changed over time. Our best chance at getting some sort of handle is through the Age at Evaluation variable. Our model (M1) indicates that the change in odds ratio of having no or mild cognitive impairment comparing second and first quantiles of education for each one unit increase in evaluation age is 1.01 (p-value > .05), and the change in odds ratio of having no or mild cognitive impairment comparing third and first quantiles of education for each one unit increase in evaluation age is 0.96 (p-value < .05). This indicates that age at evaluation is modifying the effect of education level on cognitive impairment. Further work, perhaps more longitudinally, could get a better handle on the modification of age and education.

We also see that the interaction between Residence at evaluation and Education are interacting significantly. This provides further evidence of a psycho-social association and effect on dementia and that a built environment has an association with dementia. Precisely, the odds ratio of having no or mild cognitive impairment comparing doctorate degrees and private residence to no college experience and non-private residence is 2.77. It is also a significant association (Wald's test p-value of 0.04989). While there are similar (significant) associations for all levels of education beyond high school (see Table 1 in Appendix), we cannot claim a causal relationship because it might be that higher educated Americans have access to better healthcare due to better job prospects. A better designed study, with more precise controls for socio-economic and environmental effects would allow a better causal inference.

Discussion and Conclusion

One of the problems with our data was the fact that many of the categories were imbalanced. Primarily, Education posed this problem as our predictor of interest, and we struggled to decide if we should fit it as categorical variable or continuous variable. While we ended up fitting it as a categorical variable with quartiles at cut points due to a lack of specificity and imbalance in the education levels, a fine grained resolution of education in the future would allow researchers to fit education as a continuous variable.

This study led us to several conclusions about how we would like to re-design data collection in this case. There was a dearth of non-white participants in the first dataset cleaned by Wakim, which meant that she had to categorize people into white or non-white ethnicities. This practice, while allowing us to fit models, erases the complex nature of race. It would have been better if the original sample had over-sampled minority racial and ethnic groups, particularly considering the history of how race, medicine, and power have interacted in the history of the U.S. As our final model stands, Race was insignificant, further supporting the notion that a more nuanced definition of race is necessary for questions about the association of race and dementia. Additionally, there were variables that we found in the literature review that were not included in the data set. For example, socioeconomic variables such as income, or retirement age, or occupation would be useful in deciding if there is an environmental association. Education is only a proxy for these kinds of variables, and does not contain as much information. The Education variable itself also required manipulation. Instead of being continuous, it was ordinal categorical with levels ranging from 2 to 25. The sheer number of levels indicated a need to reduce the categories. If data were somehow gathered in a continuous format, these results might be more interpretable.

Finally, as we noted above, a likelihood ratio test between a proportional odds model and a nominal odds model indicated that the proportionality assumption may have been violated by the data. But as we noted, the `vglm` command for a nominal odds model was not converging in all coefficients. Future researchers could write their own convergence code to find a way to fit a cumulative odds model that converges. Alternatively, two different binary logistic regressions could be fit; one for no dementia symptoms vs. some dementia symptoms, and one for high dementia symptoms vs. mild to no symptoms.

Contributions

Ziming Huang: Model fitting and diagnostics. **Qiaoxue Liu**: Data processing, review and editing of paper. **Abby Loe**: Wrote draft of paper for groupmates to revise, figures and tables. **Hanna Venera**: Visualizations and tables, interpretation of results, paper revisions. Everyone did equal parts of the lit review, and wrote the slides together.

References

[Ada14] Adam Simning and Yeates Conwell and Edwin van Wijngaarden. "Cognitive Impairment in Public Housing Residents Living in Western New York". In: *Social Psychiatry and Psychiatric Epidemiology* 49.3 (2014), pp. 477–485. DOI: https://dx.doi.org/10.1007\%2Fs00127-013-0712-0.

[C C18] C Chen and J M Zissimopoulos. "Racial and ethnic differences in trends in dementia prevalence and risk factors in the United States". In: *Alzheimer's & dementia* 4 (2018), pp. 510–520. DOI: https://doi.org/10.1016/j.trci.2018.08.009.

[Car10] Carol Brayne and Paul G. Ince and Hannah A. D. Keage and Ian G. McKeith and Fiona E. Matthews and Tuomo Polvikoski and Raimo Sulkava. "Education, the brain and dementia: neuroprotection or compensation?: EClipSE Collaborative Members". In: *Brain* 133.8 (July 2010), pp. 2210–2216. ISSN: 0006-8950. DOI: https://doi.org/10.1093/brain/awq185. eprint: https://academic.oup.com/brain/article-pdf/133/8/2210/16697284/awq185.pdf.

[Cat09] Catherine Helmer. "Dementia and marital status at midlife and late life". In: *BMJ* 339 (2009), b1690. DOI: https://doi.org/10.1136/bmj.b1690.

[Cle14] Clement T Loy and Peter R Schofield and Anne M Turner and John BJ Kwok. "Genetics of dementia". In: *The Lancet* 383.9919 (2014), pp. 828–840. DOI: https://doi.org/10.1016/S0140-6736(13)60630-3.

[D W91] D. W. O'Connor and P. A. Pollitt and F. P. Treasure. "The influence of education and social class on the diagnosis of dementia in a community population". In: *Psychological Medicine* 21.1 (1991), pp. 219–224. DOI: https://doi.org/10.1017/S003329170001480X.

[Dor12] Dorota Religa and Kalle Spångberg and Anders Wimo and Ann-Katrin Edlund and Bengt Winblad and Maria Eriksdotter-Jönhagen. "Dementia Diagnosis Differs in Men and Women and Depends on Age and Dementia Severity: Data from SveDem, the Swedish Dementia Quality Registry". In: *Dementia and Cognitive Disorders* 33 (2012), pp. 90–95. DOI: https://doi.org/10.1159/000337038.

[Emi11] Emily Schoenhofen Sharp and Margaret Gatz. "The Relationship between Education and Dementia An Updated Systematic Review". In: *Alzheimer Disease & Associated Disorders* 25.4 (2011), pp. 289–304. DOI: https://dx.doi.org/10.1097\%2FWAD.0b013e318211c83c.

[G A98] G Azzimondi and R D'Alessandro and G Pandolfo and F S Feruglio. "Comparative study of the prevalence of dementia in two Sicilian communities with different psychosocial backgrounds". In: *Neuroepidemiology* 17.4 (1998), pp. 199–209. DOI: https://doi.org/10.1159/000026173.

[Gen15] Geneviève Chênea and Alexa Beiser and Rhoda Aub and Sarah R. Preisc and Philip A. Wolf and Carole Dufouil and Sudha Seshadri. "Gender and incidence of dementia in the Framingham Heart Study from mid-adult life". In: *Alxheimer's and Dementia* 11 (3 2015), pp. 310–320. DOI: https://doi-org.proxy.lib.umich.edu/10.1016/j.jalz.2013.10.005.

[Gla95] Gladys Maestre and Ruth Ottman and Yaakov Stern and Barry Gurland and Michael Chun and Ming-Xin Tang and Michael Shelanski and Benjamin Tycko and Richard Mayeux. "Apolipoprotein E and Alzheimer's Disease: Ethnic Variation in Genotypic Risks". In: *Annals of Neurology* 37.2 (1995), pp. 254–259. DOI: https://doi.org/10.1002/ana.410370217.

[Mar06] Marie-Florence Shadlen and David Siscovick and Annette L Fitzpatrick and Corinne Dulberg and Lewis H Kuller and Sharon Jackson. "Education, cognitive test scores, and black-white differences in dementia risk". In: *Journal of the American Geriatrics Society* 54.6 (2006), pp. 898–905. DOI: https://doi.org/10.1111/j.1532-5415.2006.00747.x.

[Már10] Márcia L. Chaves and Ana L. Camozzato and Cristiano Köhler and Jeffrey Kaye. "Predictors of the Progression of Dementia Severity in Brazilian Patients with Alzheimer's Disease and Vascular Dementia". In: *International Journal of Alzheimer's Disease* 2010 (2010), p. 7. DOI: https://doi.org/10.4061/2010/673581.

[Pie09] Pierre Missotten and Philippe Thomas and Gilles Squelard and David Di Notte, Ovide Fontaine and Louis Paquay and Jan De Lepeleire and Frank Buntinx and Michel Ylieff. "Impact of place of residence on relationship between quality of life and cognitive decline in dementia". In: *Alzheimers Disease and Associated Disorders* 23.4 (2009), pp. 395–400. DOI: https://doi.org/10.1097/wad.0b013e3181b4cf48.

[Ric06] Richard Robinson and Teresa G. Odle. "Dementia". In: *The Gale Encyclopedia of Medicine* 2.10 (2006), pp. 1128–1131.

[Sar14] Saraa Garcia-Ptacek and Bahmana Farahmand and Ingemarc Kåreholt and Dorotae Religa and Maria Luzb Cuadrado and Maria Eriksdotter. "Mortality Risk after Dementia Diagnosis by Dementia Type and Underlying Factors: A Cohort of 15,209 Patients based on the Swedish Dementia Registry". In: *Journal of Alzheimer's Disease* 41.2 (2014), pp. 467–477. DOI: http://doi.org/10.3233/JAD-131856.

[She14] Sheryl Zimmerman and Philip D. Sloane and David Reed. "Dementia Prevalence And Care In Assisted Living". In: *Health Affairs* 33.4 (2014). DOI: https://doi.org/10.1377/hlthaff.2013.1255.

Appendix

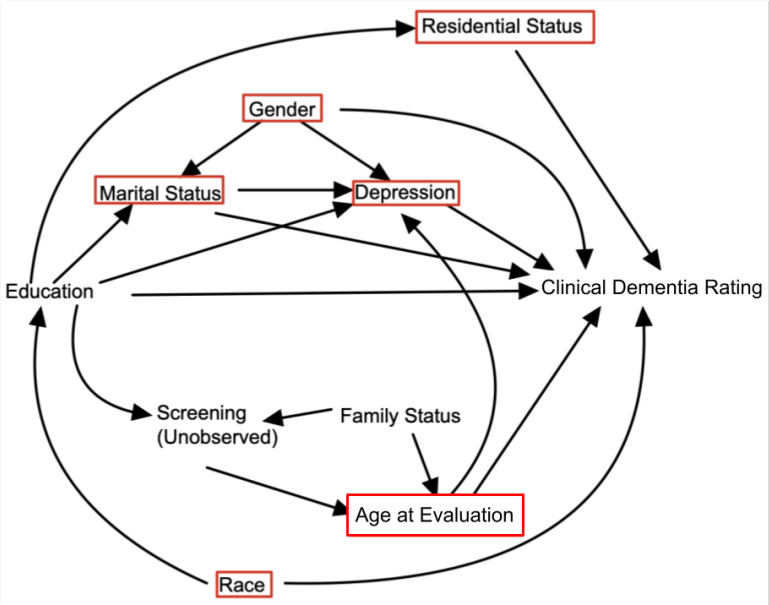

Figure 1: Relationships between observed variables

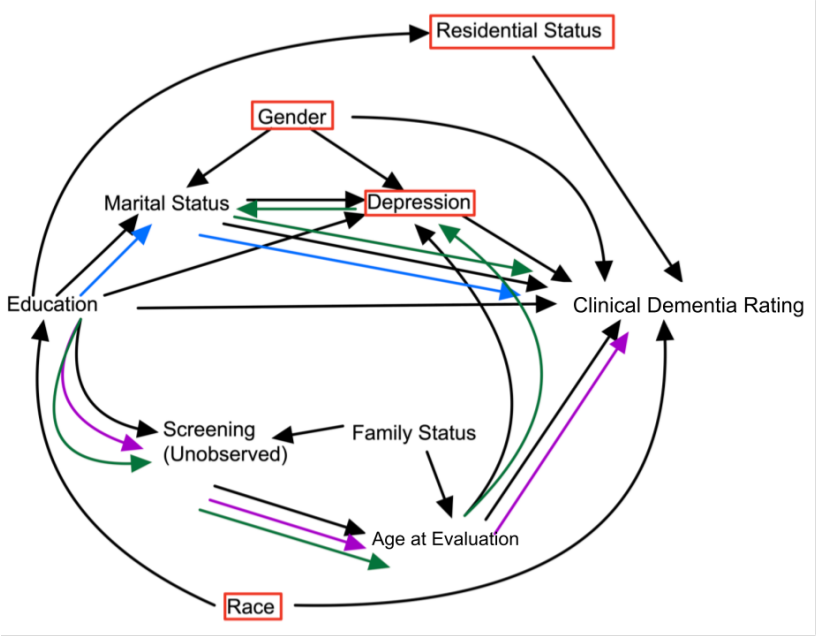

Figure 2: Relationships between observed variables

Coefficient	Estimate	Standard Deviation	P-Value
$\beta_{0\|.5}$	1.905993	0.779279	0.01445
$\beta_{.5\|1}$	3.217459	0.784323	4.09×10^{-5}
β_1	-0.666659	1.424740	0.63984
β_2	3.039931	1.184501	0.01028
β_3	-1.916881	1.578779	0.22469
β_4	-1.023344	0.117601	$< 2 \times 10^{-16}$
β_5	-0.662878	0.123047	7.16×10^{-8}
β_6	-1.330370	0.147153	$< 2 \times 10^{-16}$
β_7	-0.046109	0.009641	1.73×10^{-6}
β_8	0.116408	0.139113	0.40271
β_9	1.539297	0.226933	1.18×10^{-11}
β_{10}	0.005656	0.017847	0.75130
β_{11}	-0.038704	0.014788	0.00887
β_{12}	0.033044	0.019507	0.09028
β_{13}	1.085760	0.424597	0.01055
β_{14}	1.072787	0.354179	0.00245
β_{15}	1.019062	0.519695	0.04989

Table 1: Final Model Output

Characteristic	N = 899[1]
Education	
A	315 (35%)
B	146 (16%)
C	333 (37%)
D	105 (12%)
Sex	
Female	436 (48%)
Male	463 (52%)
Marriage status	518 (58%)
Depression	192 (21%)
First Evaluation Age	74 (69, 80)
Race	
Non-white	190 (21%)
White	709 (79%)
Residence status	659 (73%)
[1] n (%); Median (IQR)	

Figure 3: Descriptive statistics for variables in final model

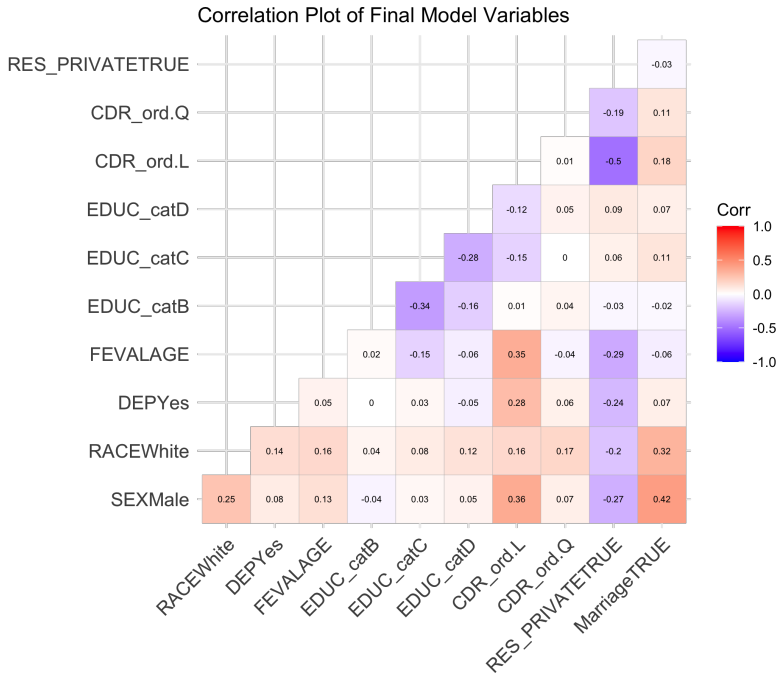

Figure 4: Correlation Plot

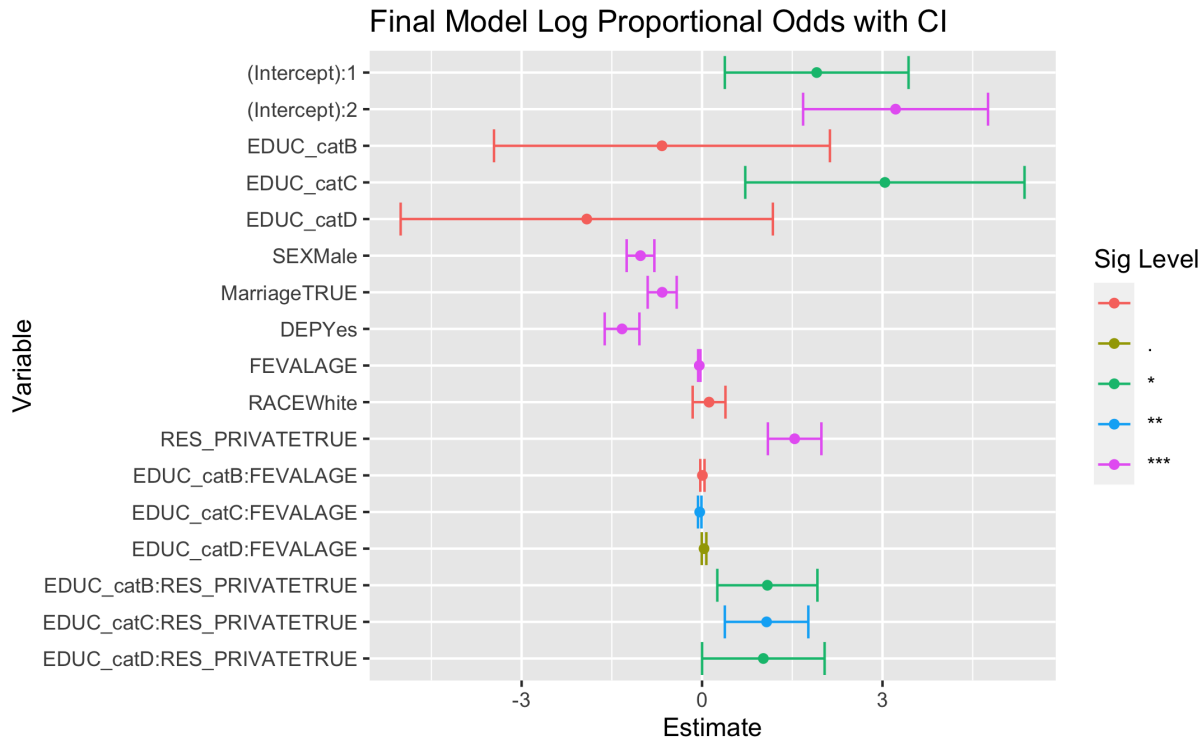

Figure 5: Final Model Results: Log Proportional Odds

g-parameter	p-value
2	0.2764082
3	0.7399551
4	0.4871606
5	0.4183525
6	0.8014770
7	0.3476331
8	0.7163588
9	0.5234130
10	0.4288111
11	0.4285228

Table 2: Results from various Homer-Lemeschow Tests

```
                GVIF Df GVIF^(1/(2*Df))
EDUC_cat    1.130946  3        1.020721
SEX         1.248450  1        1.117341
MARISTAT    1.342211  4        1.037475
DEP         1.039113  1        1.019369
FEVALAGE    1.090012  1        1.044036
RES_PRIVATE 1.036350  1        1.018013
                GVIF Df GVIF^(1/(2*Df))
EDUC_cat    1.184482  3        1.028620
SEX         1.247954  1        1.117119
MARISTAT    1.589042  4        1.059600
DEP         1.044971  1        1.022238
FEVALAGE    1.161613  1        1.077782
RES_PRIVATE 1.120105  1        1.058350
RACE        1.217719  1        1.103503
```

Figure 6: Colinearity check with gvif

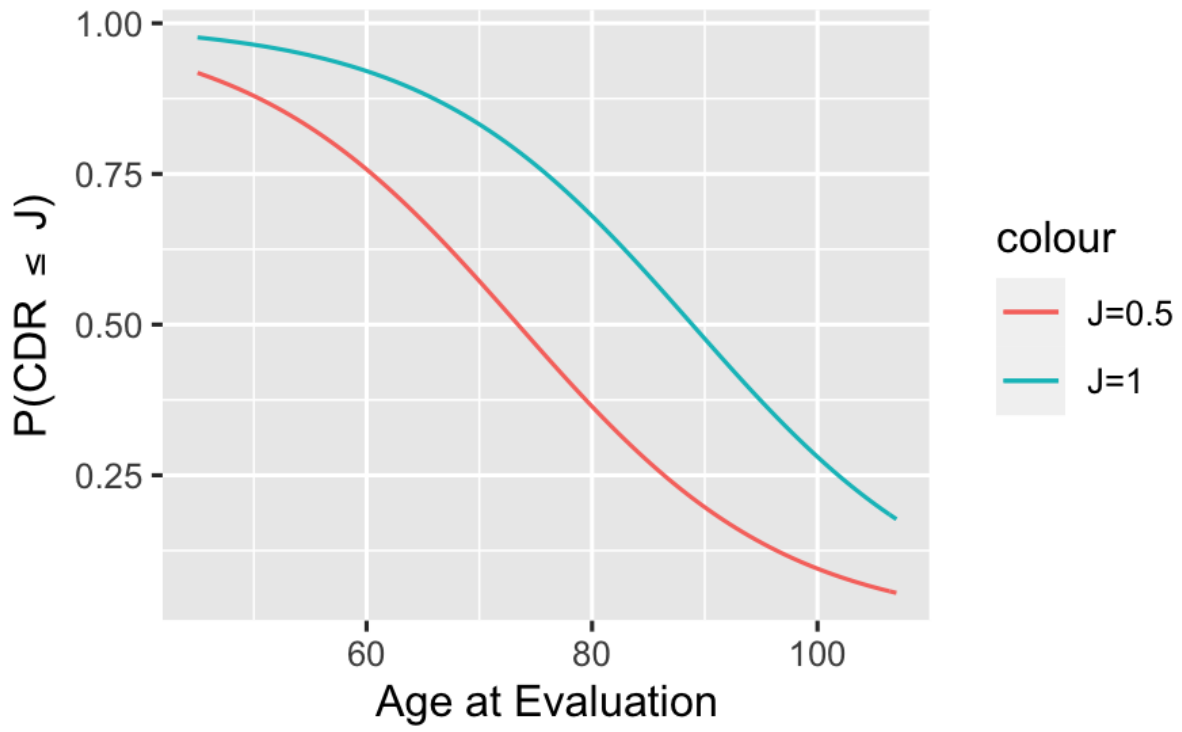

Figure 7: Model Estimates: unmarried non-white woman, with College or Master's level education, with depression, living in a private residence

```
Warning in slot(family, "validparams")(eta, y = y, extra = extra) :  It seems that the
nonparallelism assumption has resulted in intersecting linear/additive predictors.  Try
propodds() or fitting a partial nonproportional odds model or choosing some other link
function, etc.
```

Figure 8: Error thrown when fitting a nominal odds model