Ordinal Regression with Splines on Factors on Out-of-pocket Health Care Expenditure

Stat 690STA: Semiparametric Analysis

Mekhala Kumar, Vishwajth Reddy Anagandula, Ka Wing Cheung

May 24, 2023

1. Introduction

Out-of-pocket (OOP) health expenditure is closely associated with poverty. The World Health Organization (WHO) estimates that 70 million people were pushed into extreme poverty in 2017 by OOP health expenditures [1]. The association between OOP and poverty is tested using a dynamic panel threshold method with macroeconomic data from a sample of 145 countries from 2000 to 2017 [2]. Out-of-pocket health expenditure and poverty was contingent on a certain threshold level of out-of-pocket health spending [2]. A group of researchers investigates associations between socio-economic factors and out-of-pocket health spending in 16 Polish regions for the period 1999-2019 using panel-data regression [3]. Factors like disposable income, the proportions of children (aged 0-9) and elderly (70+ years) in the population, healthcare supply (proxied by physicians' density), air pollution, and tobacco and alcohol expenditure are positively associated with out-of-pocket costs [3].

In this report, we are interested in the effect of age and other predictors on OOP health expenditure along with the associations of background and insurance information of an individual. When the OOP health expenditure is a set of ordered categories, we use ordinal regression with splines terms in generalized additive model settings. In R, we have multiple

options to do that, and we will use three specific functions: *MASS::polr()*, *VGAM::vglm()*, and *mgcv::gam()*.

2. Data

We are working with the National Health Interview Survey, 2013 (ICPSR 36147) on the personal level with 104,520 rows and 610 columns [4]. The data comes from ICPSR (Inter-university Consortium for Political and Social Research), which is from the Institute for Social Research at the University of Michigan. This survey has information on all family members with respect to health status, limitation of daily activities, cognitive impairment, and health conditions, and variables related to doctor visits, hospital stays, and health care access and utilization. There are responses from children under 18 years old, but they are provided from "a knowledgeable adult", not the children themselves.

From this dataset, our response variable is *oopCost* - OOP health expenditure for the individual/ their family in the past 12 months. The final set of predictors we have chosen is a combination of demographic, health, and insurance based:

(categorical) *earning* - estimate earnings before taxes in last calendar year;

(categorical) *edu* - highest level of education completed;

(continuous) *age* - age of the person;

(categorical) *armForce* - armed force status;

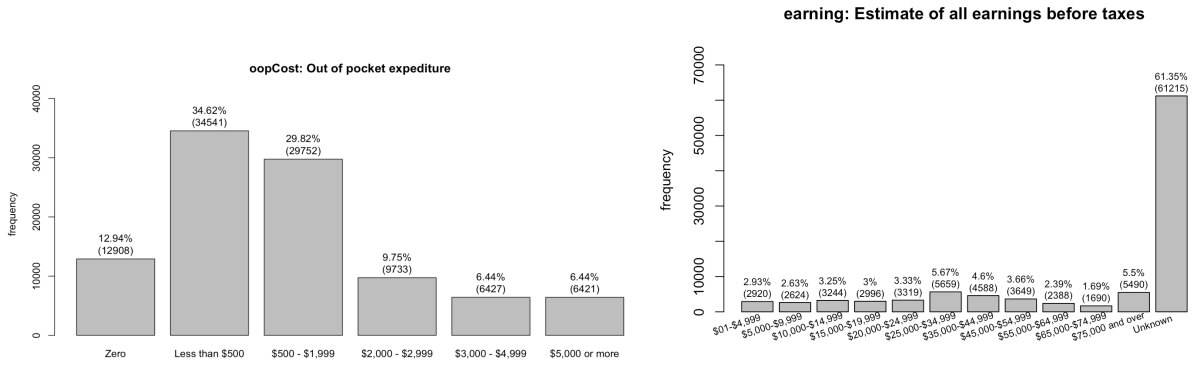
(categorical) *limitation* - any physician and mental limitation - all persons, all conditions;

(categorical) *healthStatus* - reported health status;

(categorical) *privateInsurance* - any private health insurance or not;

(categorical) *noInsurance* - no insurance coverage for any amount of time in the past 12 months.

To create useful ordinal variables for *edu* and *healthStatus*, we remove some missing values and unknown responses from the data and recode the categories: 2.88% from total for *oopCost*, 1.35% from *edu*, 0.10% from *limitation*, 0.13% from *healthStatus*, and 0.90% from *privateInsurance*. The resulting dataset has 99,782 observations from a total of 104,520 observations. The response *oopCost* has 6 ordered levels: $0 to over $5,000; *earning* has 11 levels: $0 to over $75,000 and 61.35% unknown; *edu* has 6 ordered levels: none - graduate school, *age* is continuous from 0 to 85 year old; *armForce* has 3 levels: yes, no, and 38.97% unknown, *limitation* has 2 levels: limited in any way and not limited in any way; *healthStatus* has 5 ordered levels: excellent to poor; *privateInsurance* has 2 levels: yes and no; and *noInsurance* has 3 levels: yes, no, and 15.44% unknown. The corresponding visualizations of the variables are in Figure 2.1.



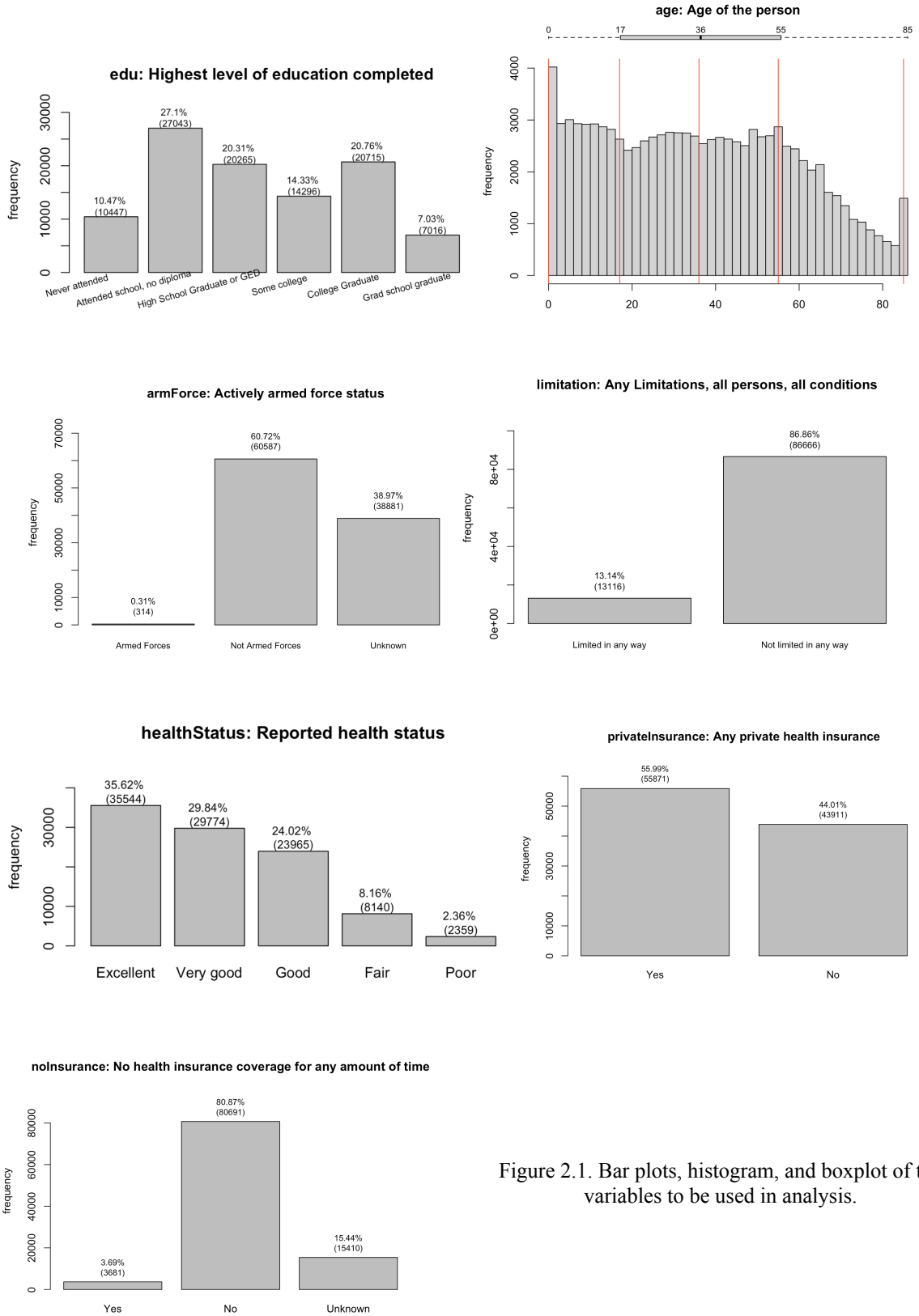


Figure 2.1. Bar plots, histogram, and boxplot of the variables to be used in analysis.

Since the variable of interest is *age*, we look at *age* versus *oopCost* in Figure 2.2. We can see that the median age as the cost of OOP health expenditure increases, but not linearly. Using spline with *age* may help with better modeling.

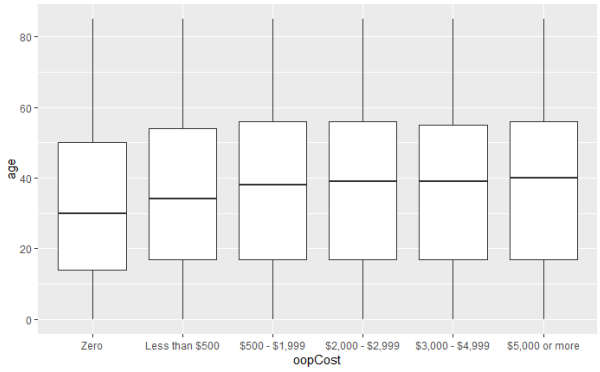


Figure 2.2. Boxplot of *age* by different levels of *oopCost*.

In this dataset, there are two identifying variables: family and household index numbers. Together, the two identifiers create a unique key that groups individuals from the same family; we call this new variable *id*. There are 40,825 unique *id* values, or unique families. This group structure suggests that we should model our data using *id* for mixed models, but the large number of values means that any modeling would be computationally expensive. Figure 2.2 shows the frequency of the number of people in each family. 62.5% of all families have less than 3 people, 90.1% have less than 5, and overwhelmingly 98.5% have less than 7 people. Given the large number of observations and families, and that a very small percentage of the families have more than 6 people, we believe that a random effect on *id* is not necessary. We will not consider modeling with random effects.

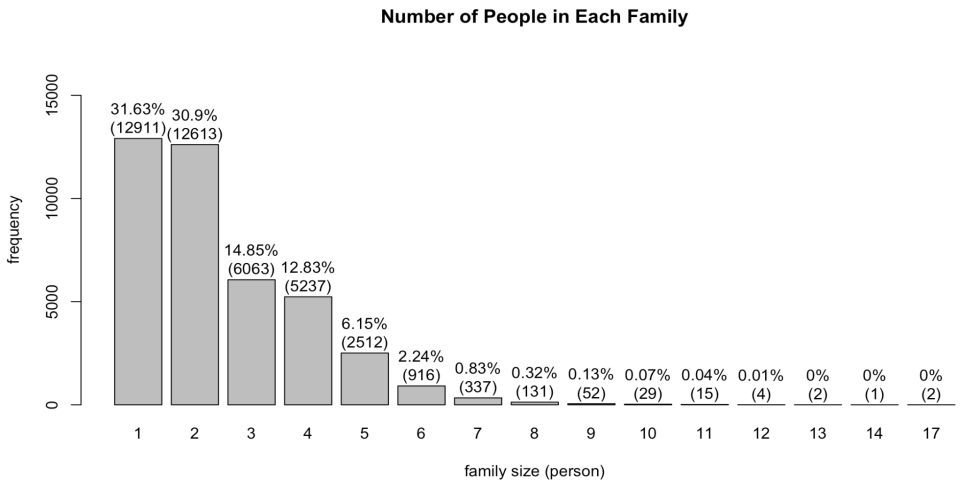


Figure 2.2. Bar plot of the number of people in each of the 40,825 families.

3. Methods

To compare the three different ordinal regression functions in R, we use a baseline model of all main effect predictors with a spline on only *age*. Then we use multiple models in each of the three different functions and compare to our baseline model in each of the three cases.

3.1 MASS::polr()

One of the models we use is the *polr()* function from the *MASS* package in R. This utilizes ordered logistic or probit regression for models with an ordered response variable. In this model, the response has to be a factor variable and can hold numeric or character values. The B-spline basis terms were added using the *bs()* function from the *splines* package in R. There is an option to include either the number of knots or degrees of freedom as the required argument in *bs()*. We tried several models with splines being fit on different variables. For the age variable, the knots were set at the quantiles and for education and health status, the knot was set at the mean. Then we used stepAIC() to select terms. Finally, we interpret the odds ratios of the coefficients.

3.2 VGAM::vglm()

We use *vglm()* function from the *VGAM* package in R, which uses vector generalized linear models and use *cumulative(parallel = TRUE)* for the family argument in vglm() for using proportional odds model, which uses an ordinal variable with more than two categories as response. The proportional odds model is described in detail in Section 3.2.1. For fitting spline terms we use *VGAM::s()* which fits a cubic smoothing spline with degrees of freedom specified as an argument. Then we interpret the fitted model for some important terms and interesting findings.

3.2.1 Proportional Odds Model

The proportional odds model is used for an ordinal response Y with m categories. In a proportional odds model $m-1$ logits are formed based on adjacent category cut points between successive categories. The cumulative log odds for category j is defined as

$$L_j = L_j(\mathbf{x}) \equiv logit[P(Y \leq j \mid \mathbf{x})] = \log \frac{P(Y \leq j \mid \mathbf{x})}{P(Y > j \mid \mathbf{x})} = \log \frac{P(Y \leq j \mid \mathbf{x})}{1 - P(Y \leq j \mid \mathbf{x})}$$

where $P(Y \leq j \mid \boldsymbol{x})$ are the cumulative probabilities given by

$$P(Y \leq j \mid \mathbf{x}) = \pi_1(\mathbf{x}) + \pi_2(\mathbf{x}) + \cdots + \pi_j(\mathbf{x})$$

where $\pi_j(\mathbf{x}) = P(Y = j \mid \boldsymbol{x})$ is the probability of response j, given the predictors $\mathbf{x}$.

Then the proportional odds model is defined as

$$L_j(\mathbf{x}) = \alpha_j + \mathbf{x}^T\boldsymbol{\beta}, \quad j = 1, \ldots, m-1$$

where α_j is an intercept for category j and $\boldsymbol{\beta}$ are coefficients of predictors. So the cumulative logits only differ in their intercepts as each category of response gets a different intercept while the effect of predictors are the same for all cumulative logits as the coefficients of the predictors stay the same for every response category j.

3.3 mgcv::gam()

One of the methods we will use is generalized additive modeling in R using the function *mgcv::gam()*. This R function has the capability to handle ordinal response and interaction terms with thin plate regression splines as the default in the smooth term *mgcv::s()*. The family for the function is *ocat(R = R)*, where its argument *R* is the number of categories in the response variable. The response variable has to also be formatted to have categories 1, 2, 3, and etc.. To perform LASSO regression on the spline terms, we will use the argument *select = TRUE*. We will fit several models and compare them using the AIC and percent of deviance explained. Then we will visualize the spline terms and interpret their effects, and visualize the prediction probabilities of the response.

4. Results

4.1 MASS::polr()

Using *MASS:polr(),* three models were fit. In all of the models, all the predictors were used and the education and health status variables were treated as continuous variables. In the first model, a spline was fit to age. In the second model, splines were fit on age, education and health status. The final model had the same spline terms as the second model, with the addition of an interaction term between the health status and earnings because although one may not be well, if they earn less, they may be hesitant to access healthcare and in contrast, if one is well but they earn more, they may be willing to spend on healthcare. One warning that occurred when the models were run was that design appeared to be rank-deficient so some coefficients were dropped. The AIC and Residual Deviance are provided in Table 4.1.1.

Model	oopCost ~	AIC	Residual Deviance
1	*bs(age,knots=quantiles(age))+earning+eduNum+armForce+limitation+ healthStatusNum+privateInsurance+noInsurance*	303,282	303,222
2	*bs(age,knots=quantiles(age))+earning+armForce+limitation+ bs(healthStatusNum,knots=mean(df$healthStatusNum))+ bs(eduNum,mean(df$eduNum))+privateInsurance+noInsurance*	303,266	303,196
3	*bs(age,knots=quantiles(age))+earning+armForce+limitation+ bs(healthStatusNum,knots=mean(df$healthStatusNum))+ bs(eduNum,mean(df$eduNum))+privateInsurance+noInsurance+ bs(healthStatusNum,knots=mean(df$healthStatusNum)):earning*	303,226	303,068

Table 4.1.1. Fitted models using MASS:polr() and their AIC and residual deviance.

Although model 3 yielded the best AIC score among the three models, due to the complexity of the model, it was not chosen as the final model. Model 2 was chosen as the final model and further interpreted.

All the terms in Model 2 were statistically significant and stepAIC() showed that all the predictors should be kept in the model. For the main independent variable age, it was found that higher chances of out-of-pocket costs occur at the knots which correspond to ages 0 and 55 respectively. With respect to earnings, the increase in out-of-pocket costs was only observed for the categories $55000- $64999, and $75000 and over. Those who are not in the Armed Forces are 1.3 times more likely to have out-of-pocket health expenditure than those who are in the Armed Forces, versus those whose status was unknown had an odds ratio of 2.19. As the self-reported health status deteriorated, the chances of having out-of-pocket costs increased. An increase in the education level, increased the chances of having out-of-pocket costs. Those without private insurance had an odds ratio of 0.29, denoting that they are less likely to have out-of-pocket costs. Similarly if one did not have insurance in the past 12 months, they were less likely to have out-of-pocket costs (odds ratio of 0.9) but for those where it was unclear if they

did not have insurance in the past 12 months, they 1.8 times more likely than those who did have insurance in the past 12 months to spend on healthcare.

```
            Zero|Less than $500    Less than $500|$500 - $1,999   $500 - $1,999|$2,000 - $2,999
                      0.1377208                      0.9682618                      4.0160134
$2,000 - $2,999|$3,000 - $4,999  $3,000 - $4,999|$5,000 or more
                      8.1587692                     17.8299747
```

Figure 4.1.1. Odds-Ratio Scale of the Response Categories

```
                                                  bs(age, knots = Xi)1                                                  bs(age, knots = Xi)2
                                                             1.2733253                                                             0.9177218
                                                  bs(age, knots = Xi)3                                                  bs(age, knots = Xi)4
                                                             0.7080090                                                             2.0400702
                                                  bs(age, knots = Xi)5                                                  bs(age, knots = Xi)6
                                                             0.5981488                                                             0.6241702
                                                 earning$5,000-$9,999                                                 earning$10,000-$14,999
                                                             0.9535371                                                             0.9265680
                                                earning$15,000-$19,999                                                earning$20,000-$24,999
                                                             0.8770851                                                             0.8850295
                                                earning$25,000-$34,999                                                earning$35,000-$44,999
                                                             0.9229607                                                             0.9298752
                                                earning$45,000-$54,999                                                earning$55,000-$64,999
                                                             0.9660676                                                             1.0355339
                                                earning$65,000-$74,999                                                earning$75,000 and over
                                                             0.9999738                                                             1.2820367
                                                      earningUnknown                                                armForceNot Armed Forces
                                                             0.9360392                                                             1.3005519
                                                    armForceUnknown                                     limitationNot limited in any way
                                                             2.1851546                                                             0.7705885
bs(healthStatusNum, knots = mean(df$healthStatusNum))1 bs(healthStatusNum, knots = mean(df$healthStatusNum))2
                                                             0.9639963                                                             1.1649886
bs(healthStatusNum, knots = mean(df$healthStatusNum))3 bs(healthStatusNum, knots = mean(df$healthStatusNum))4
                                                             1.2258182                                                             1.5963688
                                        bs(eduNum, mean(df$eduNum))1                                          bs(eduNum, mean(df$eduNum))2
                                                             1.1681703                                                             1.7317309
                                        bs(eduNum, mean(df$eduNum))3                                                 privateInsuranceNo
                                                             1.9234463                                                             0.2895101
                                                     noInsuranceNo                                                noInsuranceUnknown
                                                             0.9203961                                                             1.8144815
```

Table 4.1.2. Odds Ratios of the Predictor

Variable	p-value
bs(age, knots = Xi)1	1.15e-03
bs(age, knots = Xi)2	2.45e-01
bs(age, knots = Xi)3	9.11e-07
bs(age, knots = Xi)4	8.35e-19
bs(age, knots = Xi)5	2.32e-10
bs(age, knots = Xi)6	2.62e-11
earning$5,000-$9,999	3.37e-01

earning$10,000-$14,999	1.05e-01
earning$15,000-$19,999	6.32e-03
earning$20,000-$24,999	8.95e-03
earning$25,000-$34,999	5.75e-02
earning$35,000-$44,999	9.76e-02
earning$45,000-$54,999	4.53e-01
earning$55,000-$64,999	4.91e-01
earning$65,000-$74,999	1.00e+00
earning$75,000 and over	1.26e-08
earningUnknown	6.61e-02
armForceNot Armed Forces	1.12e-02
armForceUnknown	3.92e-13
limitationNot limited in any way	5.34e-37
bs(healthStatusNum, knots = mean(df$healthStatusNum))1	4.67e-01
bs(healthStatusNum, knots = mean(df$healthStatusNum))2	5.71e-02
bs(healthStatusNum, knots = mean(df$healthStatusNum))3	3.61e-03
bs(healthStatusNum, knots = mean(df$healthStatusNum))4	9.20e-27
bs(eduNum, mean(df$eduNum))1	5.81e-02
bs(eduNum, mean(df$eduNum))2	7.02e-30
bs(eduNum, mean(df$eduNum))3	5.63e-42
privateInsuranceNo	0.00e+00
noInsuranceNo	7.90e-03
noInsuranceUnknown	9.83e-65
Zero|Less than $500	6.20e-58
Less than $500|$500 - $1,999	7.94e-01
$500 - $1,999|$2,000 - $2,999	2.02e-29

$2,000 - $2,999|$3,000 - $4,999	1.02e-64
$3,000 - $4,999|$5,000 or more	1.2e-119

Table 4.1.3. p-values of the predictors

4.2 VGAM::vglm()

We fit three models using *VGAM::vglm()*. The first model is the full main effects model with spline fit for *age*, just like all our methods this is our baseline model. We then fit spline terms on *age*, education (*edu*), and health status (*healthStatus*) for our next model. But we find that the AIC and residual deviance of both models are the same, with AIC of 304,004 and residual deviance of 303,940. So we consider our first model, the main effects model with spline fit for age, as this model is simpler than our second model. Then we add an interaction term between private insurance (*privateInsurance*) and no insurance (*noInsurance*), as both checks for insurance and may be related, and a spline fit for health status for our third model. In this model we get a smaller AIC and smaller residual deviance than our previous two models, with AIC of 303,899 and residual deviance of 303,821. All the three models along with their AIC and residual deviance are shown in Table 4.2.1.

Model	oopCost ~	AIC	Residual Deviance
1	*earning + edu + s(age, df=25) + armForce + limitation + healthStatusNum + privateInsurance + noInsurance*	304,004	303,940
2	*earning + s(edu, df=5) + s(age, df=25) + armForce + limitation + s(healthStatus) + privateInsurance + noInsurance*	304,004	303,940
3	*earning + edu + s(age, df=25) + armForce + limitation + s(healthStatus, df=5) + privateInsurance + noInsurance + privateInsurance:noInsurance*	303,889	303,821

Table 4.2.1. Fitted models using VGAM::vglm() and their AIC and residual deviance.

The fitted coefficients of model 3 are shown in Table 4.2.2 (a). The second intercept in our model, which is the second intercept in our proportional odds model, and spline term for age is not significant. Almost all the coefficients for earnings are not significant except

$15,000-$19,999, $55,000-$64,999, and $75,000 and over when 0.05 is considered as the significance level. The rest of the coefficients are very significant except a category in education, *edu*, and a category in *noInsurance*, which are shown in Table 4.2.2 (a). All the exponentiated coefficients are given in Table 4.2.2 (b) for model 3. The cumulative odds for each response category, *oopCost*, increases as *healthStatus* increases from lowest level, Poor, to highest level, Very good, when other variables are fixed. Also the cumulative odds for each response category, *oopCost*, increases as *earnings* increase from $5,000-$9,999 to $15,000-$19,999 and then cumulative odds decreases as *earnings* increase from $15,000-$19,999 to $55,000-$64,999. Interestingly when *noInsurance* is unknown and *privateInsurance* is not present, the cumulative odds in each response category, *oopCost*, increases by 2.045 + 1.419 + 0.5279 = 3.992 times the cumulative odds in each response category when *noInsurance* is *Yes* and *privateInsurance* is also present. Also when *noInsurance* is *No* and *privateInsurance* is not present, the cumulative odds in each response category, *oopCost*, increases by 2.045 + 0.8301 + 1.780 = 4.655 times the cumulative odds in each response category when *noInsurance* is *Yes* and *privateInsurance* is present. As the residual deviance and the AIC are very high for the three models in Table 4.2.1, these models may not be a good fit for our data. The AIC and residual deviance for models 1 and 2 are the same, which implies that the splines in *VGAM::vglm()* are not smoothing enough to reduce the AIC or the residual deviance.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept):1	-1.967e+00	1.197e-01	-16.431	< 2e-16	***
(Intercept):2	-2.449e-02	1.195e-01	-0.205	0.837713	
(Intercept):3	1.391e+00	1.196e-01	11.630	< 2e-16	***
(Intercept):4	2.098e+00	1.198e-01	17.517	< 2e-16	***
(Intercept):5	2.878e+00	1.201e-01	23.968	< 2e-16	***
earning$5,000-$9,999	3.947e-02	4.881e-02	0.809	0.418742	
earning$10,000-$14,999	5.589e-02	4.637e-02	1.205	0.228119	
earning$15,000-$19,999	9.838e-02	4.732e-02	2.079	0.037624	*
earning$20,000-$24,999	8.525e-02	4.612e-02	1.849	0.064521	.
earning$25,000-$34,999	2.983e-02	4.148e-02	0.719	0.472055	
earning$35,000-$44,999	1.635e-02	4.322e-02	0.378	0.705182	
earning$45,000-$54,999	-5.156e-02	4.539e-02	-1.136	0.255933	
earning$55,000-$64,999	-1.181e-01	5.037e-02	-2.344	0.019074	*
earning$65,000-$74,999	-9.460e-02	5.574e-02	-1.697	0.089697	.
earning$75,000 and over	-3.615e-01	4.272e-02	-8.463	< 2e-16	***
earningUnknown	1.560e-03	3.524e-02	0.044	0.964699	
edu.L	-3.887e-01	2.556e-02	-15.205	< 2e-16	***
edu.Q	-7.967e-02	2.021e-02	-3.942	8.10e-05	***
edu.C	5.272e-02	1.506e-02	3.501	0.000464	***
edu^4	1.226e-02	1.335e-02	0.918	0.358514	
edu^5	-4.935e-02	1.389e-02	-3.554	0.000380	***
s(age, 25)	-2.808e-05	3.557e-04	-0.079	0.937087	
armForceNot Armed Forces	-3.268e-01	1.051e-01	-3.109	0.001880	**
armForceUnknown	-7.130e-01	1.061e-01	-6.719	1.83e-11	***
limitationNot limited in any way	2.462e-01	2.016e-02	12.214	< 2e-16	***
s(healthStatus, 5)Fair	-2.716e-01	2.534e-02	-10.719	< 2e-16	***
s(healthStatus, 5)Good	-1.528e-01	1.618e-02	-9.440	< 2e-16	***
s(healthStatus, 5)Poor	-4.879e-01	4.265e-02	-11.439	< 2e-16	***
s(healthStatus, 5)Very good	-5.007e-02	1.455e-02	-3.441	0.000579	***
privateInsuranceNo	7.153e-01	6.057e-02	11.811	< 2e-16	***
noInsuranceNo	-1.862e-01	4.071e-02	-4.573	4.82e-06	***
noInsuranceUnknown	3.501e-01	2.081e-01	1.682	0.092536	.
privateInsuranceNo:noInsuranceNo	5.764e-01	6.181e-02	9.327	< 2e-16	***
privateInsuranceNo:noInsuranceUnknown	-6.389e-01	2.135e-01	-2.992	0.002769	**

(a)

Exponentiated coefficients:

earning$5,000-$9,999	1.0402586	earning$10,000-$14,999	1.0574789
earning$15,000-$19,999	1.1033828	earning$20,000-$24,999	1.0889946
earning$25,000-$34,999	1.0302794	earning$35,000-$44,999	1.0164853
earning$45,000-$54,999	0.9497463	earning$55,000-$64,999	0.8886398
earning$65,000-$74,999	0.9097410	earning$75,000 and over	0.6966317
earningUnknown	1.0015608	edu.L	0.6779535
edu.Q	0.9234177	edu.C	1.0541314
edu^4	1.0123352	edu^5	0.9518441
s(age, 25)	0.9999719	armForceNot Armed Forces	0.7212212
armForceUnknown	0.4901879	limitationNot limited in any way	1.2791623
s(healthStatus, 5)Fair	0.7621690	s(healthStatus, 5)Good	0.8583419
s(healthStatus, 5)Poor	0.6139070	s(healthStatus, 5)Very good	0.9511671
privateInsuranceNo	2.0448461	noInsuranceNo	0.8301471
noInsuranceUnknown	1.4191593	privateInsuranceNo:noInsuranceNo	1.7796718
privateInsuranceNo:noInsuranceUnknown	0.5278538		

(b)

Table 4.2.2: (a) Fitted coefficients of model 3 and (b) exponentiated coefficients of model 3

4.3 mgcv::gam()

The models we fit with *mgcv::gam()* are presented in Table 4.3.1. Model 1 is the baseline model with a spline term for only *age*. We do want to use splines on *edu* and *healthStatus*, so we fit Model 2, with a main effect model with *eduNum* and *healthStatusNum* for spline on the two variables, along with *age*. Model 3 removes *armForce* because only 0.31% of the respondents are currently serving in the military. That could be too small for the model to detect true significance. Then we include the interaction term between *privateInsurance* and *noInsurance* in Model 4 because they are insurance related and having no insurance for any time could have an effect on having private insurance or not. Model 4 does explain more of the deviance. Then we add two-way interactions between *age* and *edu* and between *age* and *healthStatus* in Model 5. Treating the ordinal variables as numerical. This model has the smallest AIC and the largest deviance explained. LASSO variable selection on the splines shows very low p-values for all the spline terms, so we proceed with Model 4.

Model	oopCost ~	AIC	Deviance Explained
1	*earning + edu + s(age, k=25) + armForce + limitation + healthStatusNum + privateInsurance + noInsurance*	302,858	3.71%
2	*earning + s(eduNum, k=4) + s(age, k=25) + armForce + limitation + s(healthStatusNum, k=3) + privateInsurance + noInsurance*	302,869	3.7%
3	*earning + s(eduNum, k=4) + s(age, k=25) + limitation + s(healthStatusNum, k=3) + privateInsurance + noInsurance*	302,870	3.7%
4	*earning + s(eduNum, k=4) + s(age, k=25) + limitation + s(healthStatusNum, k=3) + privateInsurance + noInsurance + privateInsurance:noInsurance*	302,774	3.73%
5	*earning + s(eduNum, k=4) + s(age, k=25) + limitation + s(healthStatusNum, k=3) + privateInsurance + noInsurance + privateInsurance:noInsurance + s(age,eduNum) + s(age,healthStatusNum)*	302,639	3.79%

Table 4.3.1. Fitted models using mgcv::gam() and their AIC and percent deviance explained.

Using diagnostics with *gam.check()*, the k-index associated with the term *s(age, healthStatusNum)* is 0.97 and p-value is 0.01 with edf of 16.18. This indicate that k is too low, so we fit the model

$$oopCost \sim earning + s(eduNum, k=4) + s(age, k=25) + limitation + s(healthStatusNum, k=3) + privateInsurance + noInsurance + privateInsurance:noInsurance + s(age,eduNum) + s(age,healthStatusNum, k=35). \quad (4.3.1)$$

The corresponding AIC is 302,621 and deviance explained is 3.8%. The diagnostic plots of the model are in Figure 4.3.1. We can see that there is an outlier in the QQ plot, and there are no zero residuals. The residuals are calculated as a difference of observed and fitted values. The lack of zero indicates lack of accurate prediction. This is a sign of a bad model, since the deviance explained is also very small at 3.8%. But given the predictors we have chosen, this is the final model using *mgcv::gam()*.

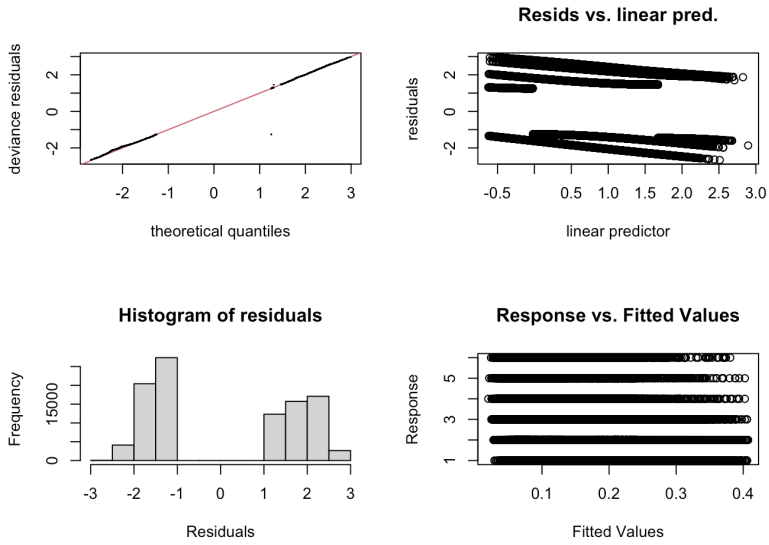


Figure 4.3.1. Diagnostic plots of Model 4.3.1.

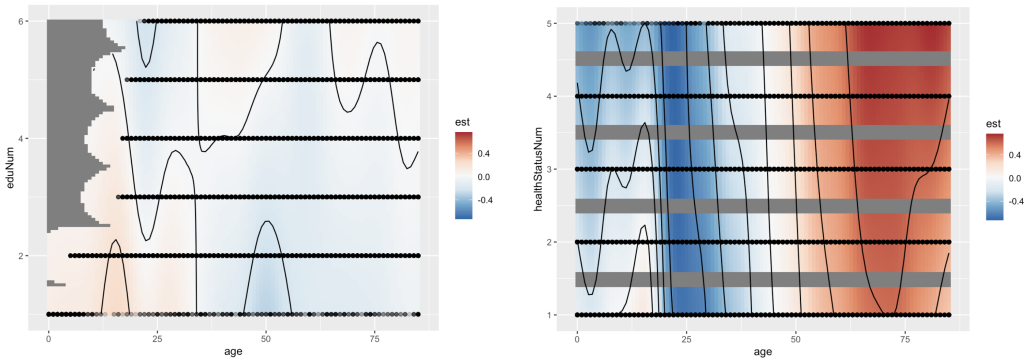
The coefficients of the model are in Table 4.3.2 where at least 1 level of all variables are significant under 0.1 with the exception of some levels of *earning* and *noInsurance* for level *unknown*. All the spline terms are significant.

```
Parametric coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               1.623299   0.055764  29.110  < 2e-16 ***
earning$5,000-$9,999                     -0.013064   0.049616  -0.263  0.79231
earning$10,000-$14,999                   -0.022729   0.047115  -0.482  0.62951
earning$15,000-$19,999                   -0.071880   0.048244  -1.490  0.13625
earning$20,000-$24,999                   -0.069364   0.047008  -1.476  0.14006
earning$25,000-$34,999                   -0.029545   0.042627  -0.693  0.48825
earning$35,000-$44,999                   -0.035510   0.044307  -0.801  0.42287
earning$45,000-$54,999                   -0.006488   0.046425  -0.140  0.88886
earning$55,000-$64,999                    0.052934   0.051111   1.036  0.30035
earning$65,000-$74,999                    0.013215   0.056251   0.235  0.81426
earning$75,000 and over                   0.248005   0.044058   5.629 1.81e-08 ***
earningUnknown                           -0.064479   0.036142  -1.784  0.07442 .
limitationNot limited in any way         -0.261412   0.020382 -12.826  < 2e-16 ***
privateInsuranceNo                       -0.747906   0.061214 -12.218  < 2e-16 ***
noInsuranceNo                             0.120606   0.041028   2.940  0.00329 **
noInsuranceUnknown                       -0.288643   0.207979  -1.388  0.16518
privateInsuranceNo:noInsuranceNo         -0.516423   0.062488  -8.264  < 2e-16 ***
privateInsuranceNo:noInsuranceUnknown     0.607189   0.213554   2.843  0.00447 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                          edf Ref.df Chi.sq  p-value
s(eduNum)               2.873  2.985 292.25  < 2e-16 ***
s(age)                  1.002  1.002  17.76 2.64e-05 ***
s(healthStatusNum)      1.588  1.827  93.60  < 2e-16 ***
s(age,eduNum)          12.645 27.000  75.41  < 2e-16 ***
s(age,healthStatusNum) 22.011 32.000 143.36  < 2e-16 ***
```

Table 4.3.2. Estimated coefficients and significance of smooth terms summary of Model 4.3.1.

The heat map and contour plot of the spline terms are in Figure 4.3.2. In (a), we have the plot of the interaction between *age* and *eduNum*. We see a negative effect on *oopCost* for young adults with higher education and older adults with less education. As age increases and education level decreases, the effect on OOP health expenditure is negative. It could be that young adults with higher education are more likely to have better insurance or have a lesser need of medical treatments, and hence less OOP cost. In (b), we have the plot of the interaction between *age* and *healthStatusNum*. We see a positive effect on *oopCost* for older people in general and young people with excellent health. Young people with reported excellent health may be more likely to get cheaper insurance and when they have medical treatments, their OOP health expenditures are higher.



(a) (b)

Figure 4.3.2. Contour plot of interaction term (a) *age:eduNum* and (b) *age:healthStatusNum*.

In ordinal regression, predictions are made for each level of the response. In Figure 4.3.3, we have 6 plots corresponding to the model's predictions for each level of *oop*Cost by their linear predictor. The linear predictors are the link scaled predictions of the ordinal regression. The corresponding range of linear predictors for each level is highlighted in red in the figure. The probability of predicting for a certain level is the highest at the corresponding range of the linear predictors. We can see that is true in the categories of "less than $500", "$500 - $1,999", and

"$2,000 - $2,999". For the other levels of *oopCost*, there are no corresponding probabilities, (linear predictors are link-scale of the probabilities). This is an indication of a poor model.

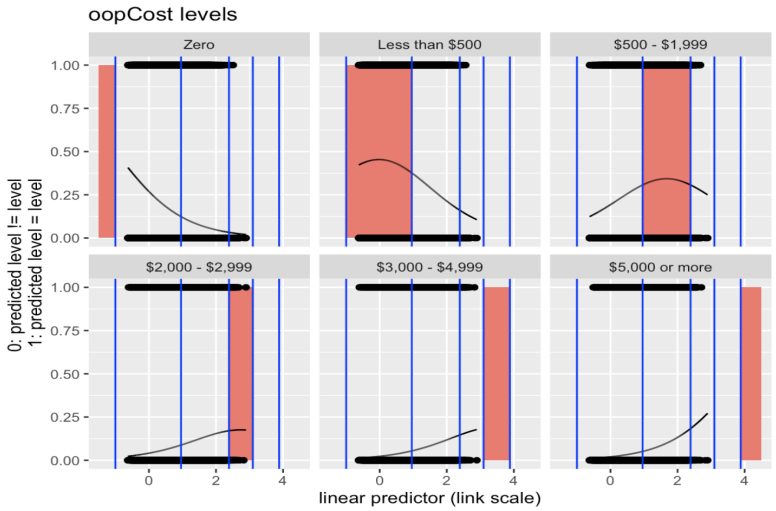


Figure 4.3.3. Prediction probabilities of *oopCost* using Model 4.3.1.
The blue lines of cutoff points of the linear predictor are -1, 0.96, 2.39, 3.1, and 3.88. The red region represents the corresponding cutoff region of the linear predictor for the *oopCost* level. The black line represents the corresponding probability based on the fitted linear predictors using the full data for each level of *oopCost*. The black points are whether the predicted level from fitting matches the observed level of *oopCost*.

5. Conclusion

	AIC	
	Final Model	**Main Effect Model with Splines only on *age***
MASS::polr()	303,266	303,282
VGAM::vglm()	303,889	304,004
mgcv::gam()	302,621	302,858

Table 5.1. AIC values of models using the three different functions.

The *MASS:polr()* final model shows higher chances of OOP health expenditure are observed at knots 1 and 4, that is ages 0 and 55 respectively. The *VGAM::vglm()* final model shows that the cumulative odds for OOP health expenditure tends to increase for all levels as health status increases from poor to very good. The *mgcv::gam()* final model shows that age has some interaction effect with education and with self reported health status. As age increases and

education level decreases, the effect on OOP health expenditure is negative. The AICs of final models with the three methods are in Table 5.1. The final model with *mgcv::gam()* has lowest AIC (see Table 5.1), but percent deviance explained is only 3.8% and there are no zero-residuals. And the prediction of certain levels of the OOP health expenditure is not good. Together, the final model of *mgcv::gam()* is not a good model, even though it is the better of the three in terms of AIC.

The function *MASS::polr()* does not have an in-built spline option. We can include B-splines or natural splines using the *bs()* and *ns()* functions in the *splines* package. However, when spline terms are introduced via these functions, the models cannot be plotted. The function *VGAM::vglm()* allows for splines but does not yet have interaction terms with spline, which is currently under development along with plotting functions. We can also use the function *vgam()*, which fits vector generalized additive models, instead of *vglm()*. The function *mgcv:gam()* allows for penalized splines, interaction terms with splines, and easy plotting of spline interaction terms. This is a more comprehensive package to use with ordinal regression and spline terms. The base model of a main effect model with splines only on *age* also from *mgcv:gam()* has the lowest AIC (see Table 5.1). Ordinal regression using the function *mgcv:gam()* is adequate and can allow for visualization of splines term when comparing with other ordinal regression functions in R, like *MASS::polr()* and *VGAM::vglm()*.

Future Directions

In future research, the relationship between age and out-of-pocket health expenditure can be studied through a Bayesian approach. The utilization of Markov chain Monte Carlo (MCMC)

samples or chains may result in a better performance of the model and explain a greater percentage of the deviance.

References

[1] World Health Organization. (2022). World health statistics 2022: monitoring health for the SDGs, sustainable development goals. World Health Organization. https://apps.who.int/iris/handle/10665/356584. License: CC BY-NC-SA 3.0 IGO

[2] Sirag A, Mohamed Nor N. Out-of-Pocket Health Expenditure and Poverty: Evidence from a Dynamic Panel Threshold Analysis. Healthcare (Basel). 2021 May 3;9(5):536. doi: 10.3390/healthcare9050536. PMID: 34063652; PMCID: PMC8147610.

[3] Łyszczarz B, Abdi Z. Factors Associated with Out-of-Pocket Health Expenditure in Polish Regions. Healthcare (Basel). 2021 Dec 17;9(12):1750. doi: 10.3390/healthcare9121750. PMID: 34946475; PMCID: PMC8701368.

[4] United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. National Health Interview Survey, 2013. Inter-university Consortium for Political and Social Research [distributor], 2015, https://doi.org10.3886/ICPSR36147.v1.