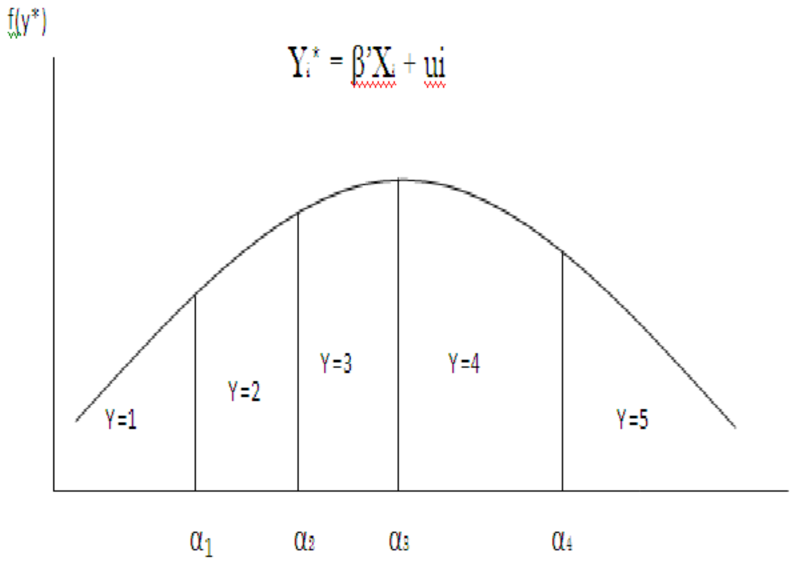
ORDINAL AND SEQUENTIAL DISCRETE CHOICE MODEL

Copyright of Abhinav Anand, Jyoti Arora and Shraddha Ramswamy

April 19th 2013

INTRODUCTION
Researchers in Health Economics have long been interested in the utility of perceived health as an indicator of health status in Health Economics. Many studies of self-rated health show that it is a reliable predictor of health status even when controlling for health-related variables and status characteristics. According to previous research, one reason for the consistent finding is that self-ratings of health represent judgements of health trajectories.
This paper investigates the impact of a host of personal and status characteristics such as age, level of education, race and residence in Southern or Northern region (w.r.t Baseline) on how the citizens of United States perceive their health for the year 1992 using ordinal and sequential logistic model.
DATA
The dataset is taken from NHANES Epidemiological Follow Up Study:1992 wave.
Age is measured in years, education is measured in terms of number of years of schooling completed and dichotomous variable is created for gender (female = 1) and race (black = 1).
METHODOLOGY
We use ordinal and sequential logistic model respectively for assessing the impact of personal characteristics like age and education and status characteristics like Southern residence and race on self perception of health status.
1. **ORDINAL LOGIT**
In an ordered model, the response Y (here the self rated health) is restricted to one of m ordered value (here from 1 to 5). The cumulative logit model assumes that the ordinal nature of the observed response is due to methodological limitations in collecting the data that results in lumping together values of an otherwise continuous response. Here, the self rated health measure which is the dependent variable takes value from 1 to 5. It is assumed that the unobservable variable (i.e the self perception of



health status) is a continuous latent variable Y* such that:

$$Y = i \text{ when } \alpha(i-1) < Y* < \alpha(i) \text{ where i= 1,2,3,4,5}$$
$$-\propto = \alpha_0 < \alpha(1) < \alpha(2) < \alpha(3) < \alpha(4) < \alpha(5) = \propto .$$

It is further assumed that the latent variable Y* is determined by the explanatory variable vector X (consisting of age, schooling, race and gender) in the linear form $Y* = \beta' X_i + u$ where β is vector of coefficients; and u is random variable with distribution function described by F( ).
It follows that :

$$P(y_i = j) = P(\alpha_{j-1} \leq y_i* \leq \alpha_j)$$
$$P(y_i = j) = P(\alpha_{j-1} \leq \beta' x_i + u_i \leq \alpha_j)$$
$$P(y_i = j) = P(\alpha_{j-1} - \beta' x_i \leq u_i \leq \alpha_j - \beta' x_i)$$
$$P(y_i = j) = F(\alpha_j - \beta' x_i) - F(\alpha_{j-1} - \beta' x_i)$$

Where j = 1, 2, 3, 4, 5 and i is the ith individual

Since U follows a logistic distribution function, the cumulative model is also called the proportional odds model. Since u has a logistic distribution,

$$F(U_i) = \frac{e^{U_i}}{1+e^{U_i}}$$
$$f(u_i) = \frac{e^{U_i}}{1+e^{U_i}}^2$$
$$P(y_i = j/x_i) = \frac{e^{\alpha_j - \beta' x_i}}{1+e^{\alpha_j - \beta' x_i}} - \frac{e^{\alpha_{j-1} - \beta' x_i}}{1+e^{\alpha_{j-1} - \beta' x_i}}$$

Where j = 1, 2, 3, 4, 5 and i represents the ith individual

$$Y_i = 1 \text{ then } P_{i1} = F[\alpha_1 - \beta' X_i]$$
$$Y_i = 2 \text{ then } P_{i2} = F[\alpha_2 - \beta' X_i] - F[\alpha_1 - \beta' X_i]$$
$$Y_i = 3 \text{ then } P_{i3} = F[\alpha_3 - \beta' X_i] - F[\alpha_2 - \beta' X_i]$$
$$Y_i = 4 \text{ then} P_{i4} = F[\alpha_4 - \beta' X_i] - F[\alpha_3 - \beta' X_i]$$
$$Y_i = 5 \text{ then} P_{i5} = 1 - F[\alpha_4 - \beta' X_i]$$

where F() is defined as above.
For estimating the model we specify 5 dummy variables for the i^{th} individual with
the following rule:

$$Z_{ij} = 1 \text{ if } Y_i = j \text{ where j = 1,2,3,4,5. } Z_{ij} = 0 \text{ otherwise.}$$

Then, assuming U as logistic distribution f(Ui),

$$L_i \equiv \prod_{j=1}^{5} P_{ij}^{z_{ij}} = \prod_{j=1}^{5} [\frac{e^{\alpha_j - \beta' x_i}}{1+e^{\alpha_j - \beta' x_i}} \text{ - } \frac{e^{\alpha_{j-1} - \beta' x_i}}{1+e^{\alpha_{j-1} - \beta' x_i}}]^{z_{ij}}$$

As the observations are independent, the likelihood function is product of individual likelihood functions:

$$L \equiv \prod_{i=1}^{3712} \prod_{j=1}^{5} P_{ij}^{z_{ij}}$$
$$\equiv \prod_{i=1}^{3712} \prod_{j=1}^{5} [\frac{e^{\alpha_j - \beta' x_i}}{1+e^{\alpha_j - \beta' x_i}} - \frac{e^{\alpha_{j-1} - \beta' x_i}}{1+e^{\alpha_{j-1} - \beta' x_i}}]^{z_{ij}}$$

Since likelihood functions are globally concave, we use Newton Raphson method to compute β.

$$\hat{\beta}_j = \hat{\beta}_{j-1} - [\frac{\partial^2 LogL}{\partial \beta^2}]^{-1} * [\frac{\partial LogL}{\partial \beta}] \mid \hat{\beta}_{j-1}$$

RESULTS
For ordered logit regression, the following command was used in SAS :

```
proc logistic data = sasuser.nhanes descending;
model health = age gender race edu south;
run;
```

and the results obtained were as follows :

INFERENCE

The LOGISTIC Procedure

Model Information

Data Set	SASUSER.NHANES	
Response Variable	health	health
Number of Response Levels	5	
Model	cumulative logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	3712
Number of Observations Used	3712

Response Profile

Ordered Value	health	Total Frequency
1	5	699
2	4	1141
3	3	1088
4	2	556
5	1	228

Probabilities modeled are cumulated over the lower Ordered Values.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
37.0697	15	0.0012

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	11088.082	10550.236
SC	11112.960	10606.210
-2 Log L	11080.082	10532.236

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	547.8468	5	<.0001
Score	498.1829	5	<.0001
Wald	532.3508	5	<.0001

Analysis of Maximum Likelihood Estimates

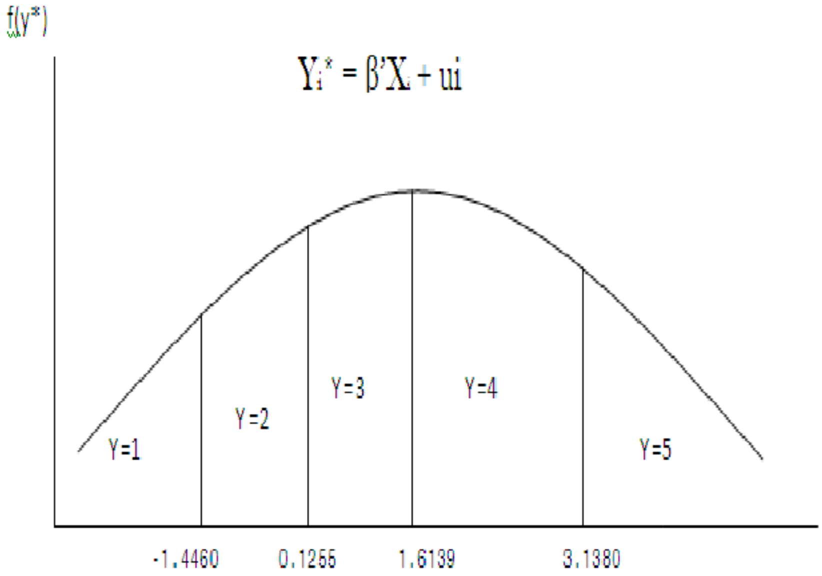
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept 5	1	-1.4460	0.2473	34.1904	<.0001
Intercept 4	1	0.1255	0.2463	0.2598	0.6103
Intercept 3	1	1.6139	0.2479	42.3953	<.0001
Intercept 2	1	3.1380	0.2539	152.7003	<.0001
Age	1	-0.0313	0.00262	143.3251	<.0001
gender	1	0.00989	0.0605	0.0267	0.8701
race	1	-0.2122	0.0669	10.0676	0.0015
edu	1	0.1553	0.0114	184.0970	<.0001
south	1	-0.7989	0.1072	55.5218	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.969	0.964	0.974
gender	1.010	0.897	1.137
race	0.809	0.709	0.922
edu	1.168	1.142	1.194
south	0.450	0.365	0.555

Association of Predicted Probabilities and Observed Responses

Percent Concordant	65.8	Somers' D	0.322
Percent Discordant	33.6	Gamma	0.324
Percent Tied	0.6	Tau-a	0.244
Pairs	5221799	c	0.661



- **Intercept Parameters**
  The intercept parameters represent the thresholds of the choices. These can be represented as follows:

- **Slope parameter for age**
  One additional year of age results in a 3.13 percent decreases in odds ratio of health being self rated as

  - excellent than as good,
  - very good than as good ,
  - good than as fair and
  - fair than as poor

  controlling for gender, education, race and southern residence at baseline.

- **Slope parameter for gender**
  There is almost negligible difference for females over males in the odds of rating their health as excellent than very good, or very good than good, or good than fair, or fair than poor, controlling for age, education, race and southern residence at baseline.

- **Slope parameter for race**

Blacks are 19.12 percent less likely than whites to self rate their health as

- excellent than as good,
- very good than as good ,
- good than as fair and
- fair than as poor,

controlling for age, gender, education, and southern residence at baseline.

- **Slope parameter for education**
  An additional year of schooling leads to 16.80 percent increase in odds ratio of health being self rated as

  - excellent than as good,
  - very good than as good ,
  - good than as fair and
  - fair than as poor,

  controlling for age, gender, race and southern residence at baseline.

- **Slope parameter for southern residence at baseline**
  The Southern residents in each district are 55 percent less likely than the northern residents to self rate their health status as

  - excellent than as good,
  - very good than as good ,
  - good than as fair and
  - fair than as poor,

  controlling for age, gender, race and education.

- **Concordance and Discordance**
  A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value.
  If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is discordant. If a pair of observations with different responses is neither concordant nor discordant, it is a tie. In our model, 65.8 percent of the total pairs are concordant while

33.6 percent are discordant and 0.6 percent of the total pairs form a tie which is a robust result.

2. SEQUENTIAL LOGIT

We want to analyse the the factors that explain the health perception of US Citizens using Sequential Logit Model. Assume that there are five possible levels of self-rated health. Let Y_i represent the self-rated level of the individual i. Then Y_i can take one of the four values described below:

$Y_i = 1$ if the individual i rates as "Poor"
$Y_i = 2$ if individual i rates as "Fair"
$Y_i = 3$ if individual i rates as "Good"
$Y_i = 4$ if individual i rates as "Very Good"
$Y_i = 5$ if individual i rates as "Excellent"

Let $P_{ij} = P(y_i = j|X_i)$ where $i = 1, 2, 3, ...3712$ and $j = 1, 2, 3, 4, 5$.

Then the probabilities can be written as,

$$P_{i1} = F(\beta_1' X_i)$$
$$P_{i2} = [1 - F(\beta_1' X_i)][F(\beta_2' X_i)]$$
$$P_{i3} = [1 - F(\beta_1' X_i)][1 - F(\beta_2' X_i)][F(\beta_3' X_i)]$$
$$P_{i4} = [1 - F(\beta_1' X_i)][1 - F(\beta_2' X_i)][1 - F(\beta_3' X_i)][F(\beta_4' X_i)]$$
$$P_{i5} = [1 - F(\beta_1' X_i)][1 - F(\beta_2' X_i)][1 - F(\beta_3' X_i)][1 - F(\beta_4' X_i)]$$

Observations . Five choices , and hence we have 4 latent variables to describe choices. Choices in each step are independent of the previous step.

For example,

$$P(y_i = 3) = P[Y_i \neq 1 \text{ and } Y_i \neq 2 \text{ and } Y_i = 3|Y_i \neq 1 \text{ and } Y_i \neq 2]$$
$$P(y_i = 3) = P[Y_i \neq 1 \text{ and } Y_i \neq 2]P[Y_i = 3|Y_i \neq 1 \text{ and } Y_i \neq 2]$$
$$P(y_i = 3) = P[Y_i \neq 1]P[Y_i \neq 2]P[Y_i = 3|Y_i \neq 1 \text{ and } Y_i \neq 2]$$

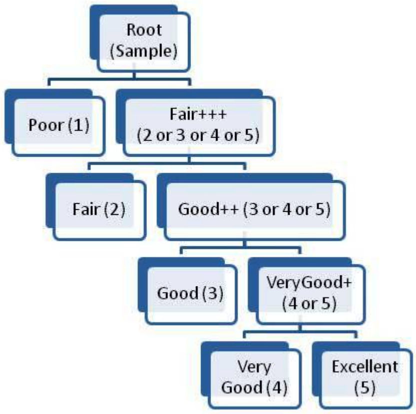
Estimation:

$$L_i = \prod_{j=1}^{5} P_{ij}^{z_{ij}}$$

Independent examples implies,

$$L = \prod_{i=1}^{3712} \prod_{j=1}^{5} P_{ij}^{z_{ij}}$$

$$\text{logL} = \sum_{i=1}^{3712} \sum_{j=1}^{5} Z_{ij} \log P_{ij}$$



Decision-theoretic Tree Structure to depict Sequential Logit Model

However, in sequential choice models the log likelihood function can be maximized by repeatedly maximizing the log likelihood functions of the associated binary models.

For sequential logistic regression, the following program was executed in SAS :

```
data seqlogit;
set seqlogit;
fairplus = (shm > 1);
fair = (shm=2);
if fairplus = 1;
run;
proc format;
value shm 1='poor' 2-5='fair+++';
value gender 0='male' 1='female';
value race 0='white' 1='black';
value resid 0='north' 1='south';
run;
proc qlim data=seqlogit; *covest=qml;
class race resid gender;
endogenous fair  discrete(dist=logistic order=formatted);
model fair = age gender race edu resid;
format gender gender. race race. resid resid.;
run;
```

		The QLIM Procedure			
		Parameter Estimates			
Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-0.9028	0.40898	-2.21	0.0273
Age*		0.031085	0.004264	7.29	<.0001
Gender	female	-0.03239	0.098606	-0.33	0.7426
Gender	male	0	.	.	.
Race	black	0.12122	0.10717	1.13	0.258
Race	white	0	.	.	.
Edu*		-0.15498	0.018192	-8.52	<.0001
Resid*	south	-1.03592	0.142367	-7.28	<.0001
Resid	north	0	.	.	.

CONCLUSION

From the above results we see that as age increases by an additional year, people generally decrease their rating of health. This result coincides with intuition as people get older their health condition deteriorates.

No difference is observed between females and males over rating their health status.

Blacks generally rate their health lower than that done by whites.This may be attributed to the discrimination that blacks face in accessing health services as opposed to whites.

With an increase in years at school, people generally rate their health more highly. This can be explained as through education, people become more aware about health related issues and services available, and thus can avoid many illnesses.

Southern residents generally rate their health status lower than the residents living in northern part of the same district. This may be because southern residents usually have lower access to health facilities than their northern counterparts .