2 September 2016

Choice Defines Value: a predictive modeling competition in health preference research

Michał Jakubczyk, PhD
Decision Analysis and Support Unit, SGH Warsaw School of Economics, Warsaw, Poland
Benjamin M. Craig, PhD*
Department of Economics, University of South Florida, Tampa, Florida, USA
Mathias Barra, PhD
Health Services Research Centre, Akershus University Hospital, Lørenskog, Norway
Catharina G. M. Groothuis-Oudshoorn
Department of Health Technology and Services Research, University of Twente, Twente, The Netherlands
John D. Hartman, MA
Haas Center for Business Research and Economic Development, University of West Florida, Pensacola, Florida, USA
Elisabeth Huynh, PhD
Institute for Choice, University of South Australia, Adelaide, Australia
Juan M. Ramos-Goñi, MSc.
EuroQol Research Foundation, Rotterdam, The Netherlands
Elly A. Stolk, PhD
EuroQol Research Foundation, Rotterdam, The Netherlands
Kim Rand-Hendriksen, PhD
Department of Health Management and Health Economics, University of Oslo, Oslo, Norway; Health Services Research Centre, Akershus University Hospital, Lørenskog, Norway

Funding/Support: Funding support for this research was provided by the EuroQol Research Foundation and a grant from the National Institutes of Health, Department of Health and Human Services, through the National Cancer Institute (1R01CA160104). The funding agreements ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. The views expressed by the authors in the publication do not necessarily reflect the views of the EuroQol Group.

Conflicts of interest: There are no conflicts of interest.
Running Title: Choice Defines Value
Word Count: 3212
Online Appendices: Screenshots, Team Submissions, Predictions
Intended Journal: Journal of Health Economics

Address correspondence to: Benjamin M. Craig, PhD, Department of Economics, University of South Florida, 4202 E. Fowler Avenue, CMC206A, Tampa, FL 33620, Phone: (813) 974-4252, Fax: (813) 974-6510, email: bcraig@usf.edu

The sections in blue will be written by Michał Jakubczyk after the announcement at the 2016 EuroQol Plenary Meeting in Berlin, Germary on 16 September 2016.

ABSTRACT

Background: Health is not bought and sold openly. Therefore, in order to understand the value of various attributes thereof, health preference researchers conduct surveys using such elicitation techniques as paired comparisons (asking, for example, “Which do you prefer?”). In the absence of external data with which to validate their findings, researchers may estimate models that are suboptimal in terms of predictive validity. Instead of conducting analyses that rely heavily on conventional practices and assumptions, Drs. Craig and Rand-Hendriksen proposed a predictive modeling competition.

Aims: To engage a large group of scientists in a modeling competition in order to identify specifications and approaches to model selection that better predict health preferences.

Methods: Announced in March 2016, the competition attracted 18 teams from around the world. In April, an exploratory survey was fielded in response to which 4074 US respondents completed 20 out of 1560 paired comparisons by choosing between two health descriptions (e.g., between longer lifespan and better health). The paired comparison data were promptly distributed to all 18 teams and were also posted online. By early July, ten teams had dropped out of the competition and eight teams had submitted their predictions for 1600 additional pairs and described their analytical approach to modeling health preferences. After these predictions had been posted online, a confirmatory study was fielded that included the 1600 additional pairs and 4148 additional respondents.

Results: In September, the victorious team, “Discreetly Charming Econometricians” led by Michał Jakubczyk, was announced at the EuroQol Plenary in Berlin, Germany. This team achieved the smallest chi square 4391.54, and their predictions were rejected by only 134 of the 1600 pairs (8%) in the confirmatory survey at a p-value of 0.01. (MORE)

Conclusions: This crowdsourcing endeavor demonstrated the diversity of analytical approaches and highlighted the importance of predictive validity in health preference research. (MORE) Although the generalizability to specific applications is unknown, the models that are more predictive may inform and inspire debate among health preference researchers.

Crowdsourcing is the process of obtaining services, ideas, or content by soliciting contributions from a large group of people rather than by relying on a single person or a handful of collaborators. By gathering the ideas of multiple independent teams, such a communal endeavor fosters greater creativity and tends to achieve a wider range of possible solutions and perspectives. This predictive modeling competition was designed on the premise that the community of health preference researchers is diverse in modeling expertise and perspectives.(Craig et al., 2016) Instead of relying on convention, peer review, or theoretical assumptions, the competition produced a diversity of analytical approaches.

Health preference research (HPR) is a scientific enterprise: specifications are devised, hypothesized, and tested. Its mantra is "choice defines value." However, by convention HPR typically conducts just one study, estimates just one analytic specification, and promotes implementation without confirmation. It seems misguided to ground health policy decisions on preliminary evidence acquired and presented from the perspective of a single team. More troubling is the approach of researchers who estimate multiple specifications and cherry-pick their results (as in data mining).(Mannila, 1997) In clinical trials, analysis plans must be formally registered before collecting and examining the data, and the results are typically confirmed by multiple teams before putting them into practice. In addition to demonstrating a diversity of analytical approaches, this competition was designed to promote scientific rigor in HPR by having multiple teams compete in their pursuit of predictive validity.

To the best of our knowledge, this is the first predictive modeling competition in HPR. Predictive validity is particularly important in HPR, because health is not bought and sold openly. Therefore, in order to understand the value of various attributes thereof, health preference researchers conduct surveys using such elicitation techniques as paired comparisons.(Berg, 1972) For the predictive modeling competition reported here, data on paired comparisons from an exploratory survey were distributed to multiple teams so that each team might apply their own modeling specifications independently. Using their findings, each team submitted predictions for a second, confirmatory set of paired comparisons. After their predictions had been posted publicly, a confirmatory survey was fielded and the teams' submissions were ranked in accordance with their predictive validity (smallest to largest chi square). Although the competition has only one winner, this crowdsourcing endeavor was also designed to yield benefits more generally to the HPR community: by promoting greater understanding of the merits underlying alternative modeling specifications, promoting the importance of predictive validity in HPR, and demonstrating the diversity of analytical approaches among HPR researchers.

METHODS

Team registration

In March 2016, Dr. Craig and Rand-Hendriksen distributed an announcement inviting all interested teams to participate in the predictive modeling competition.(Craig and Rand-Hendriksen, 2016) By April, 18 teams had registered using a brief form on the IAHPR website (no exclusions) that asked five questions pertaining to (1) conditional agreement for teams; (2) the names of the team and team leader, and the number of co-investigators; (3) the names of the co-investigators; (4) experience; and (5) invoicing. By May, all registered teams had received the exploratory data and a sample submission. By July, eight of the 18 teams had submitted their forms and predictions, and were paid $2500 (see Appendix). In September, the victorious team received a small trophy at 2016 EuroQol Plenary in Berlin and lead authorship of this manuscript in concordance with the Vancouver criteria.(2010)

Task and Pair Selection

The design of the paired comparisons (e.g., Figure 1) was largely based on the recent protocols for the Valuation Technology developed by the EuroQol Group (EQ-VT).(Oppe et al., 2016) The wording differed from the EQ-VT in four ways. (1) Because it was designed to elicit preferences, not judgments, the competition survey instrument asked "Which do you prefer?" instead of "Which is better?" (2) The labels "A" and "B" were dropped, since they might imply rank. (3) The differentiating attributes and numbers were bolded. (4) Each description included the timing and duration of problems (e.g., "Starting today, [X] years with health problems: [health state] then die ([X] years from today)").

The set of 1560 pairs in the exploratory survey was based on the 196 pairs in the EQ-VT. Every pair had two health descriptions, each of which included five problems based on the EQ-5D-5L (Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression). Each problem was characterized as being at one of five possible levels (none [level 1], slight, moderate, severe, unable/extreme [level 5]). As a shorthand notation, the five problems can be characterized as a vector of five numbers (e.g., Figure 1 includes 33333). The problems based on the 196 pairs of EQ-VT (and four ancillary pairs) had durations in four different temporal units (days, weeks, months, years), creating 800 efficient pairs.

In addition to the 800 efficient pairs, 760 time trade-off (TTO) pairs of identical structure were included. In TTO pairs, however, one health description involves no health problems (i.e., 11111) and a longer lifespan (e.g., Figure 1), like a conventional TTO task. To select the TTO pairs, 38 descriptions were selected from the efficient pairs, included durations in four different temporal units, and paired with five durations with no health problems (38×4×5=760).

The set of 1600 pairs in the confirmatory survey included 800 efficient pairs and 800 TTO pairs and were created using a similar process, albeit with some important differences. We began by selecting health descriptions that commonly occur in clinical data. The motivation for emphasis on prevalent outcomes is that health preference estimates are commonly applied to summarize health outcomes captured in clinical trials as a means to better inform medical recommendations and resource allocation decisions (e.g., cost-utility analyses). After combining the problems in these prevalent descriptions with durations to create a candidate set, we applied a software package (NGENE) to select a subset set of 200 pairs by d-efficiency. (ChoiceMetrics, 2012; Norman, 2016) To select the TTO pairs, 40 descriptions were selected from the efficient pairs. Each of the efficient and TTO pairs was shown with four temporal units, creating 1600 confirmatory pairs. A full description of the process of pair selection was included in the rules of the competition and distributed in advance of team registration.(Craig and Rand-Hendriksen, 2016) All pairs are shown in the Appendix.

Health Preference Survey

Between March 21 and April 6 (exploratory) and between July 25 and August 26, 2016 (confirmatory), 8,222 US adults aged 18 and older were recruited from a nationally representative panel to participate in a 25-minute online survey. The survey instrument had four components (see Appendix): Screener, Health, Paired Comparison, and Follow-up. The Screener captured the respondent's consent and demographic and socioeconomic characteristics (Table 1). Respondents who passed the Screener (i.e., who consented and did so within demographic quotas) were asked to complete the Health Component, including a five-level question on general health, the EQ-5D-5L and visual-analogue scale (range of worst to best, 0 to 100) on general health. After viewing three examples of paired comparisons, each respondent completed 10 efficient pairs followed by 10 TTO pairs. In the Follow-up Component, respondents were asked "How would you describe this survey?" and offered eight adjectival statements (Table 2) shown in random order with three response options (Not True, Sometimes True, Often True).

Aside from their fielding dates, the only differences between the exploratory and confirmatory surveys were the pairs. The order of the pairs was randomized at the respondent level, and the two alternatives within each pair were randomized horizontally at the respondent level (i.e., left and right) such that the shorter lifespan was either always on the left or always on the right. Each of the 3200 pairs had approximately 50 responses following 18 demographic quotas (all combinations of two genders, three age groups, and three race or ethnicity groups) to promote concordance with the 2010 US Census. Screenshots of the survey instrument is provided in the Appendix.

Econometric Analysis

To aid in the interpretation of the results, the respondents are described in terms of their demographic and socioeconomic characteristics. We assessed for differential dropout and differences between the exploratory and confirmatory surveys using chi square tests. Respondents are also described in terms of their response behaviors (e.g., always choosing the left option), lexicographic patterns in the paired-comparison responses (e.g., always choosing longer lifespan), and their follow-up descriptions of the survey. Regardless of how they responded to the survey, all respondents were included in the analytical sample.

To simplify the analysis, each team received only the paired-comparison responses of the exploratory survey, not also the respondent characteristics. As part of the submission form, the teams were explicitly asked whether this exclusion made the modeling more difficult (See Appendix). Using these preference data, the teams independently estimated their models and submitted predictions for the pairs in the exploratory and confirmatory surveys, ranging from 0.000 to 1.000 (see Appendix).

To characterize fit of each team's predictions, we computed the chi square:

$$\sum N_k \times (y_k - p_k)^2 \times (y_k \times (1-y_k))^{-1}$$

In this formula, N_k is the sample size (e.g., 50 responses), p_k is the team's prediction, and y_k is the sample probability for the k^{th} pair. If a sample probability was unanimous (y_k =1 or 0), the weight, $(y_k \times (1-y_k))^{-1}$, was replaced with the Berkson weight, $(4N_k^3)/(2\times N_k-1)$.(Berkson, 1955) Although the team with the smallest chi square based on the confirmatory survey wins the competition, the chi square of the confirmatory survey (y-axis) is plotted against that of the exploratory survey as an indicator of differential fit.

To illustrate which predictions are rejected by the data at a p-value of 0.01, an immediate form of the binomial test was run for each pair and team prediction. For the victorious team, this concordance between the predictions and the confirmatory responses is shown in a scatterplot in which the red dots represent rejections (Figure 3). When the confirmatory responses on any pair reject a prediction, this suggests poor predictive validity (Table 3). We also calculated Lin's concordance between predictions and sample probabilities as an absolute measure of concordance (Lawrence, 1989) and explored which specific pairs were the most difficult to predict across all teams (as shown by the number of rejections).

To facilitate comparison of fit across pair subsets, reduced chi square (i.e., chi square divided by the number of pairs within a pair subset) was estimated by team, survey (exploratory or confirmatory), temporal unit, difference in survival (immediate death, half or less, or more than half) and pair type (efficient or TTO).(Andrae et al., 2010)

RESULTS

Among the 13,974 US adults recruited by email for the study (i.e., survey visits), 12,123 (87%) completed the screener, 9,212 (66%) were selected to participate in the study (i.e., consented and were quota sampled), 8,721 (62%) completed the health component, and 8,222 (59%) completed the paired comparisons (median 18.26 minutes; 13.52-27.02 IQR). The 990 respondents who dropped out during the Health or Paired Comparison Components were often female, Black or African American, or reported "Refused/Don't know" for household income than respondents who completed the Components (Table 1; p-value<0.01). Although there were no significant differences in respondent demographics, the exploratory and confirmatory samples have small differences in education attainment and household income (<5%). The analytical samples has higher educational attainment and household income compared to the US 2010 Census.

In Table 2, we show patterns in the response behaviors (p-value<0.01). Few respondents chose only left or only right (<0.5%). Some respondents always chose the shorter lifespan (<2%) and others (5%) always chose the alternative with the longer lifespan. These behaviors and lexicographic preferences were slightly less prevalent in the exploratory survey compared to the confirmatory survey (<3% difference). Based on the reported descriptions of the survey, most considered the survey to be "Interesting, thought provoking, eye opening" (90%) and "Challenging, tricky, tough and difficult" (78%). Less than half considered the survey "Ridiculous, implausible, unrealistic," "Enjoyable, amusing, entertaining, fun," or "Unclear, vague, and nebulous." The descriptions of the exploratory and confirmatory surveys were similarly, except that a few additional respondents (<3% difference) indicated that the confirmatory survey was "Weird, unusual, bizarre, odd, strange."

Figure 2 shows the chi square for the confirmatory and exploratory surveys by team. Among the eight teams, chi square ranged from 908.78 to 5587.42 for the exploratory survey and from 4391.54 to 8028.86 for the confirmatory survey. Among the eight teams, "Discreetly Charming Econometricians led by Michał Jakubczyk, submitted the predictions with the lowest chi square of the confirmatory survey (4391.54). Their predictions also had the fewest number of rejections and the highest concordance as measured by Lin's rho, suggesting that their analytical approach clearly had the greatest predictive validity.

Table 3 shows the reduced chi square by team, survey and pair types. All teams had the most difficulty predicting preferences in "years" and the least difficulty doing so in "weeks." The teams' predictive validity differed greatly for dead pairs (pairs including "immediate death") and were more similar for lifespan pairs (i.e., where a shorter lifespan was paired with a longer lifespan). All teams predicted the TTO pairs better than the efficient pairs in the exploratory survey, but this was not the case for the confirmatory survey.

Comparison of Analytical Approaches

The submission forms enabled teams to describe the process or rationale according to which they selected their model (See Appendix); and this, as we see it, may be of greater importance than the models and estimation techniques. Four of the eight teams built from the example (Fedora), which was a Bradley-Terry model with a power function to relax the constant proportionality assumption. Instead of characterizing all of the models in greater detail here, we refer the reader to the team's descriptions of their analytical approaches.

In the confirmatory survey, the four teams that used the Bradley-Terry model (Occam's Barbershop Quartet, Fedora, Marginal Choices and Preferential Treatment) performed worse on predicting the dead pairs, but largely better on predicting the lifespan pairs compared to the other four teams, which suggests that the Bradley-Terry model (or its power function) may need to be modified for the

prediction of dead pairs. Likewise, the four teams that used the Bradley-Terry model largely predicted the efficient pairs better and the TTO pairs worse than their competitors. Based on these results, the best analytical approach may be to estimate a Bradley-Terry model to predict the efficient pairs and an alternative approach to predict the dead and TTO pairs.

(MORE ON THE VICTORIOUS MODEL)

DISCUSSION

The purpose of this study was not to conclude that a specific model should be promoted as universally "best," or as "true" in some deeper sense. The main objective was to use crowdsourcing to get as many model specifications as possible out in the open so that their strengths and weaknesses could be discussed (See Appendix). The competition successfully promoted the importance of predictive validity in HPR, and showed the diversity of analytical approaches and perspectives within the HPR community. This competition has drawn attention to critical problems with specific econometric issues and modeling approaches (e.g., logits, constant proportionality) as well as disseminated a public dataset and code so that student and scientists who are new to the field have a better understanding of the challenges of health preference modeling.

(MORE ON SUMMARY OF THE FORMS)

This project faced some challenges related to an unexpected aspect of the competition design. All 18 of the registered teams openly agreed with the amount of compensation ($2500), the rules of the competition, and the time frame for its deliverables. Chi square was selected from the set of all possible valid measures of fit as the primary means of assessing predictive validity for the competition. Knowing that they would be ranked by their chi square, each team had an incentive to submit values that minimized predictive error on the pair probabilities. Nevertheless, ten of the 18 teams dropped out after receiving the exploratory data, because (1) their intended approach performed poorly (e.g., logits); (2) the attributes involved unexpected complexities common to health valuation (e.g., different temporal units); or (3) team leaders had to attend to unexpected personal or work commitments.

When the competition was announced, some researchers expressed concern about the inherent advantages of teams involved in administering the competition. To avoid potential conflicts of interest, Dr. Craig distributed his submission (i.e., form, predictions, code) to all teams before accepting any other submissions. His submission served as an example and allowed him to review the submissions of others without provoking concern that he might then modify his own. But his example may, in turn, have contributed to the decisions of some teams to drop out and may also have induced the unintended consequence that four of the eight submissions applied a similar analytical approach to modeling, reducing analytic diversity.

Among the ten teams that dropped out, some researchers who specialize in identifying subgroups or individuals with distinct preferences (i.e., preference heterogeneity) expressed deep reservations after seeing the data, stating that predictive validity is an inherently flawed objective. From their perspective, preference data must be individually cleansed of respondent behaviors and traits before they can be properly interpreted as preferences. If one assumes that preferences are inherently latent, the prediction of confirmatory preference data is critically confounded by underlying unobservable factors. Furthermore, the selection of the chi square as the measure of fit was arbitrary (as any other measure would be) (Canary et al., 2016; Hosmer et al., 1997) and an added source of debate. The fact that ten teams dropped out after agreeing to the rules and examining the data exemplifies the diversity of analytical approaches and perspectives in HPR as well as a limitation of this competition.

Apart from its latency, we further recognize that preferences may be heterogeneous and that this competition tells us little about predicting preferences of specific individuals or differences in individual perceptions.(Craig et al., 2015) The teams predicted preferences of the general population in aggregate; however, this evidence does not imply that their models predict preferences at the individual level. Models that perform well in aggregate data may perform poorly when predicting the preferences of a specific individual, and vice versa. Individual preferences may be lexicographic (e.g., perfect complements under Leontief utility); (Leontief, 1966) and, therefore, may be poorly expressed as a continuous function. The literature on ecological fallacies includes many examples where markets act predictably but individuals do not.(Loney and Nagelkerke, 2014; Stigler, 1950) To demonstrate that a model is a suitable for individual prediction would require a respondent-specific analysis, which is planned for a future competition.

Also, it is important to acknowledge the potential biases in panel-based surveys, which is particularly challenging for experimental studies. (Craig et al., 2013) In this study, low socio-economic status (SES) is rare in online panels and associated with dropping out and with non-trading behavior (e.g., always choosing the alternative on the right or with the longest lifespan). Lexicographic response patterns may be attributable to preferences, inattentiveness, or greater cognitive difficulty. Therefore, even if online panels were able to recruit a sufficient number of participants with low SES (external validity), the responses may not reflect their actual preferences (internal validity). Furthermore, it is reasonable to expect that the exploratory and confirmatory data may differ due to seasonal or other unobservable changes in the panel. These limitations should be balanced against the feasibility of controlling such biases and its potential implication for the competition results.

(CONCLUSION)

REFERENCES

ClinicalTrials.gov[Internet]. National Library of Medicine (US): Bethesda, Maryland, USA.

Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication. Journal of Pharmacology & Pharmacotherapeutics 2010;1; 42-58.

Andrae R, Schulze-Hartung T, Melchior P. 2010. Dos and don'ts of reduced chi-squared. Cornell University Library; 2010.

Berg RL. Health Status Indexes: Proceedings of a Conference Conducted by Health Services Research, Tucson, Arizona, October 1-4, 1972s. Hospital Research and Educational Trust: Tucson, Arizona; 1972.

Berkson J. Maximum Likelihood and Minimum Chi Square Estimates of the Logistic Function. Journal of the American Statistical Association 1955;50; 130-162.

Canary JD, Blizzard L, Barry RP, Hosmer DW, Quinn SJ. Summary goodness-of-fit statistics for binary generalized linear models with noncanonical link functions. Biometrical Journal 2016;58; 674-690.

ChoiceMetrics. Ngene 1.1.1 User Manual & Reference Guides.: Australia; 2012.

Craig BM, Hays RD, Pickard AS, Cella D, Revicki DA, Reeve BB. Comparison of US panel vendors for online surveys. Journal of medical Internet research 2013;15; e260.

Craig BM, Mühlbacher A, Lancsar E, Brown DS, Ostermann J. Health Preference Research: An Overview. Medical Care 2016;Under review.

Craig BM, Pickard AS, Rand-Hendriksen K. Do health preferences contradict ordering of EQ-5D labels? Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation 2015;24; 1759-1765.

Craig BM, Rand-Hendriksen K. 2016. EQ DCE Competition Description, Rules, and Procedures v1.1. International Academy of Health Preference Research: Tampa, Florida, USA; 2016.

Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine 1997;16; 965-980.

Lawrence IKL. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 1989;45; 255-268.

Leontief W. Input-output economics. Oxford University Press: New York,; 1966.

Loney T, Nagelkerke NJ. The individualistic fallacy, ecological studies and instrumental variables: a causal interpretation. Emerging Themes in Epidemiology 2014;11; 1-6.

Mannila H. 1997. Methods and Problems in Data Mining. Proceedings of the 6th International Conference on Database Theory. Springer-Verlag; 1997.

Norman R. 2016. Appendix on Ngene Pair Selection. International Academy of Health Preference Research: Tampa, Florida, USA; 2016.

Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goñi JM, Luo N. EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. PharmacoEconomics 2016; 1-12.

Stigler GJ. The Development of Utility Theory. II. Journal of Political Economy 1950;58; 373-396.

Figure 1. Example of a Paired Comparison

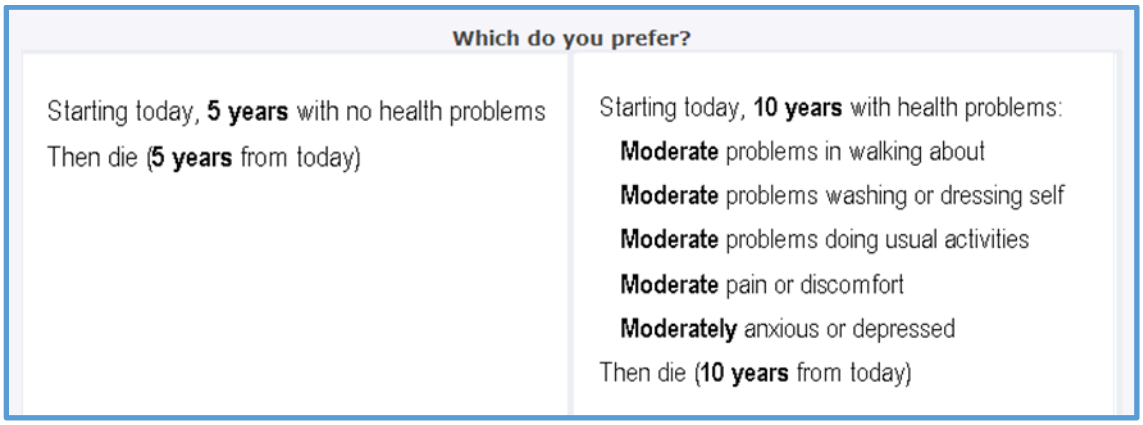


Figure 2. Exploratory and Confirmatory Chi-square by Team

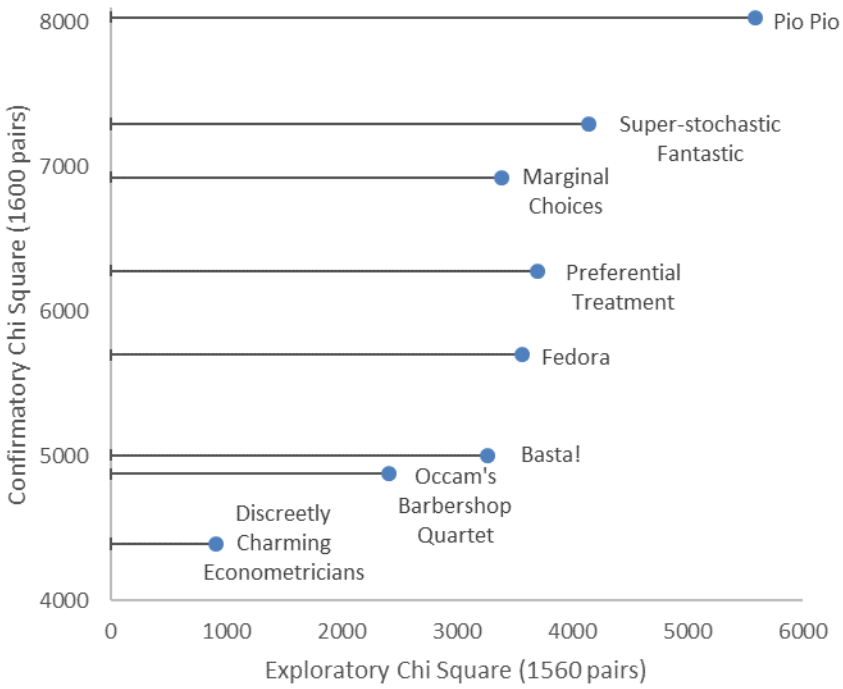


Figure 3. Predictions and Confirmatory Results for “Discreetly Charming Econometricians”

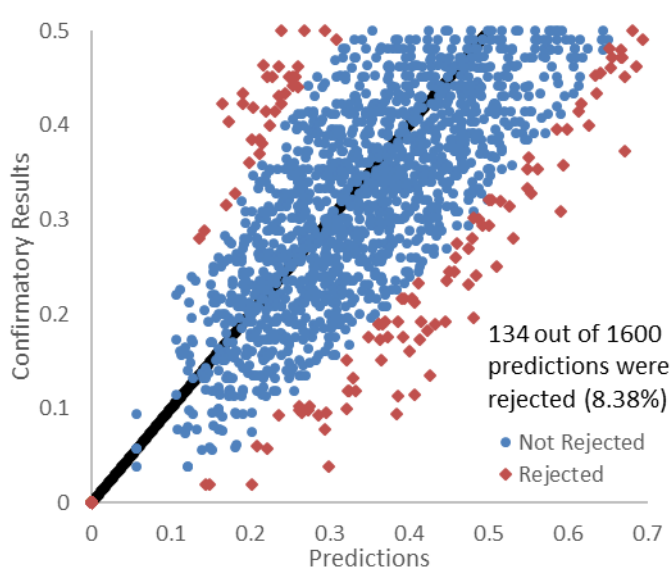


Table 1. Respondent Characteristics by Completion and Compared to 2010 US Population*

		Complete		Dropout versus	Exploratory versus	
	Dropout N=990 % (#)	Exploratory N=4074 % (#)	Confirmatory N=4148 % (#)	Complete p-value	Confirmatory p-value	US 2010 Census %
Age in years						
18 to 34	25.66 (254)	27.12 (1105)	28.21 (1170)	0.22	0.55	30.58
35 to 54	38.59 (382)	36.25 (1477)	35.68 (1480)			36.70
55 and older	35.76 (354)	36.62 (1492)	36.11 (1498)			32.72
Sex						
Male	42.83 (424)	49.39 (2012)	50.53 (2096)	<0.01	0.30	48.53
Female	57.17 (566)	50.61 (2062)	49.47 (2052)			51.47
Race						
White	77.58 (768)	82.18 (3348)	80.54 (3341)	0.02	0.11	74.66
Black or African American	16.36 (162)	11.73 (478)	12.73 (528)			11.97
American Indian or Alaska Native	0.51 (5)	0.56 (23)	0.48 (20)			0.87
Asian	2.83 (28)	2.82 (115)	3.01 (125)			4.87
Native Hawaiian or other Pacific Islander	0.40 (4)	0.59 (24)	0.39 (16)			0.16
Some other race	2.32 (23)	2.11 (86)	2.84 (118)			5.39
Two or more races						2.06
Hispanic ethnicity						
Hispanic or Latino	12.02 (119)	12.03 (490)	12.30 (510)	0.90	0.71	14.22
Not Hispanic or Latino	87.98 (871)	87.97 (3584)	87.70 (3638)			85.78
Educational attainment among age 25 or older						
Less than high school	2.42 (24)	1.91 (78)	2.10 (87)	0.96	<0.01	14.42
High school graduate	43.84 (434)	43.27 (1763)	45.20 (1875)			28.50
Some college, no degree	11.11 (110)	12.20 (497)	8.87 (368)			21.28
Associate's degree	6.97 (69)	7.44 (303)	5.67 (235)			7.61
Bachelor's degree	29.19 (289)	27.91 (1137)	30.30 (1257)			17.74
Graduate or professional degree	3.43 (34)	3.61 (147)	3.71 (154)			10.44
Refused/Don't know	0.10 (1)	0.07 (3)	0.07 (3)			-
Household Income						
$14,999 or less	5.35 (53)	4.52 (184)	4.44 (184)	<0.01	<0.01	13.46
$15,000 to $24,999	7.07 (70)	5.65 (230)	5.74 (238)			11.49
$25,000 to $34,999	8.59 (85)	8.27 (337)	8.20 (340)			10.76
$35,000 to $49,999	14.75 (146)	15.34 (625)	12.85 (533)			14.24
$50,000 to $74,999	20.51 (203)	21.23 (865)	21.79 (904)			18.28
$75,000 to $99,999	12.83 (127)	15.56 (634)	14.05 (583)			11.81
$100,000 to $149,999	12.63 (125)	13.21 (538)	15.41 (639)			11.82
$150,000 or more	6.67 (66)	7.51 (306)	9.81 (407)			8.14
Refused/Don't know	11.62 (115)	8.71 (355)	7.71 (320)			-

* Age, sex, race, and ethnicity estimates for the US are based on 2010 Census Summary File 1. Educational attainment and household income are based on 2010 American Community Survey 1-Year Estimates. Unlike the US Census, the American Community Survey excluded adults not in the community (e.g., institutionalized) and describes income by the proportion of households, not adults.

Table 2. Response Behavior, Lexicographic Preferences and Survey Description

	Exploratory N=4074 % (#)	Confirmatory N=4148 % (#)	p-value
Response Behavior			
Always Left or Always Right	0.34 (14)	0.48 (20)	0.33
Both	99.66 (4060)	99.52 (4128)	
Lexicographic Preference			
Always Shorter Lifespan	0.69 (28)	1.66 (69)	<0.01
Both	95.14 (3876)	91.88 (3811)	
Always Longer Lifespan	4.17 (170)	6.46 (268)	
Survey Description (ranked by frequency)			
Interesting, thought provoking, eye-opening			
Not True	10.75 (438)	9.81 (407)	0.30
Sometimes True	44.11 (1797)	44.31 (1838)	
Often True	43.47 (1771)	44.67 (1853)	
Challenging, tricky, tough, difficult			
Not True	22.21 (905)	23.41 (971)	0.51
Sometimes True	50.93 (2075)	50.36 (2089)	
Often True	25.14 (1024)	25.12 (1042)	
Weird, unusual, bizarre, odd, strange			
Not True	29.06 (1184)	27.12 (1125)	<0.01
Sometimes True	44.99 (1833)	44.17 (1832)	
Often True	23.93 (975)	27.34 (1134)	
Depressing, sad, scary, distressing			
Not True	30.83 (1256)	29.39 (1219)	0.03
Sometimes True	45.97 (1873)	45.40 (1883)	
Often True	21.33 (869)	23.79 (987)	
Morbid, morose, dismal, bleak, grim, somber			
Not True	31.81 (1296)	29.77 (1235)	0.02
Sometimes True	44.48 (1812)	44.62 (1851)	
Often True	21.89 (892)	24.37 (1011)	
Ridiculous, implausible, unrealistic			
Not True	53.17 (2166)	51.95 (2155)	0.35
Sometimes True	35.13 (1431)	36.31 (1506)	
Often True	9.89 (403)	10.51 (436)	
Enjoyable, amusing, entertaining, fun			
Not True	56.70 (2310)	55.91 (2319)	0.56
Sometimes True	31.76 (1294)	32.98 (1368)	
Often True	9.72 (396)	9.88 (410)	
Unclear, vague, nebulous			
Not True	56.77 (2313)	57.52 (2386)	0.93
Sometimes True	34.46 (1404)	34.28 (1422)	
Often True	6.92 (282)	6.92 (287)	

Table 3. Predictive validity, rejected predictions and reduced chi square by temporal unit, lifespan and pair type

	Predictive Validity		Rejected Predictions*	Reduced Chi Square**								
				Temporal Unit				Ratio of lifespans			Pair Type	
	Chi square	Lin's Rho	%	Days	Weeks	Months	Years	Immediate death	Half or less	More than half	Efficient	TTO***
Exploratory (N=4074)												
Discreetly Charming Econometricians	908.78	0.98	0.06	0.54	0.46	0.71	0.62	0.14	0.50	0.73	0.73	0.44
Occam's Barbershop Quartet	2415.13	0.94	3.40	1.57	1.44	1.53	1.66	1.03	1.43	1.75	1.78	1.32
Basta!	3267.00	0.93	5.96	2.06	1.90	2.23	2.19	0.81	1.79	2.62	2.67	1.52
Fedora	3569.92	0.92	4.17	2.22	1.84	2.56	2.53	1.01	1.93	2.89	3.18	1.36
Preferential Treatment	3704.80	0.91	6.99	2.80	1.81	1.88	3.01	5.67	2.31	2.25	2.45	2.11
Marginal Choices	3391.08	0.92	3.33	2.14	1.74	2.44	2.38	1.49	1.82	2.72	2.94	1.36
Super-stochastic Fantastic	4150.17	0.92	6.22	2.60	2.33	3.18	2.54	0.58	2.24	3.40	3.88	1.42
Pio Pio	5587.42	0.89	9.36	3.22	2.59	4.77	3.75	0.79	2.74	4.96	5.38	1.74
Confirmatory (N=4148)												
Discreetly Charming Econometricians	4391.54	0.87	8.38	2.63	2.42	2.64	3.30	5.13	2.35	3.13	3.29	2.04
Occam's Barbershop Quartet	4874.75	0.85	10.25	2.76	2.66	2.71	4.06	5.05	2.66	3.43	2.81	3.19
Basta!	5005.13	0.84	12.13	3.15	2.88	2.90	3.58	2.14	3.10	3.22	3.72	2.55
Fedora	5697.52	0.82	11.44	3.15	3.11	3.00	4.98	6.14	3.13	3.97	3.12	3.89
Preferential Treatment	6279.42	0.82	13.63	2.09	2.41	3.15	8.04	14.43	3.75	3.53	3.11	4.23
Marginal Choices	6924.78	0.78	15.25	3.85	3.40	3.89	6.17	9.70	3.81	4.69	4.32	4.05
Super-stochastic Fantastic	7292.12	0.78	14.69	4.67	3.47	4.65	5.44	3.52	4.40	4.82	5.28	3.86
Pio Pio	8028.86	0.77	18.44	4.08	3.12	4.45	8.43	1.85	4.80	5.49	6.21	3.93

* Rejected prediction is the proportion of pairs, where the team's prediction was rejected by the data at a p-value of 0.01 based on an immediate form of the binomial test (e.g., red dots in Figure 3).

** Reduced chi square is the chi square divided by the number of degrees of freedom (a.k.a., mean square weighted deviation). For this table, we divided by the number of pairs; therefore, reduced chi square may be interpreted as the mean of weighted squared error across the pairs.

*** The TTO pairs excludes those pairs including "immediate death," which are shown in the 8th column, "Immediate death."

3 September 2016

EQ DCE Predictive Modeling Competition

Team Submission Forms

Benjamin M. Craig, Kim Rand-Hendriksen

To facilitate the comparison of modeling approaches, each team submitted responses to 10 questions on model description, modeling recommendations and competition recommendations. Incomplete forms or forms with partial/unclear responses were returned. The responses were arranged in a common format, including the team logo. Although the scientific content was not changed, all forms were sent to an external proofreading service to make corrections regarding grammar and punctuation. Prior to posting, we sent the edited forms and combined prediction file to all teams and gave them one week to submit minor corrections. Afterwards, the forms and predictions were posted at iahpr.org.

Questions

Model Description:

1. Describe your choice of software and the reasons underlying your choice (e.g., Stata)

2. Describe your choice of estimation technique and the reasons underlying your choice (e.g., Bayesian)

3. Describe your choice of functional form and the reasons underlying your choice (e.g., Logit)

4. Describe your choice of variables and the reasons underlying your choice (e.g., 20 effects-coded variables)

Modeling Recommendations:

5. Did you have difficulty modeling the 2 pair types (TTO pairs [quantity vs. quality] and efficient pairs [all attributes])? Did you have difficulty with the 4 temporal units (days, weeks, months, years)?

6. Do you believe that you would have been able to predict choice probabilities better had you received data on the respondent characteristics as part of the exploratory dataset (e.g., age)? Why?

7. Did you change your model's functional form or variables based on the estimation results (i.e., data mining)? If so, why and how? If not, why not?

8. If your model wins, why do you believe it predicted better than the other models? If your model loses, why do you believe it did not predict better than the other models?

9. Based on your expertise and experience, what are the primary econometric advances needed to improve predictive modeling (not design)?

Competition Recommendations:

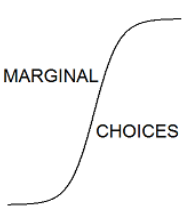
10. What recommendations do you have to improve the competition?

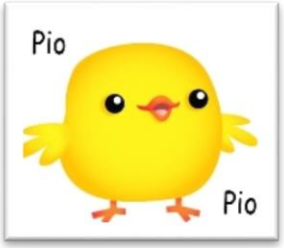
Teams (Shown in Order of Predictive Validity)

	Team Name	Team Leader	Team Members
1	Discreetly Charming Econometricians	Michał Kosma Jakubczyk	Bogumił Kamiński, Dominik Golicki**, Michał Lewandowski, Beata Koń, Paweł Ekk-Cierniakowski
2	Occam's Barbershop Quartet	Kim Rand-Hendriksen**	Mathias Barra, Liv Ariane Augestad**, Fredrik Dahl
3	Basta!	Mathias Barra	Liv Ariane Augestad**
4	Fedora	Benjamin M. Craig*,**	
5	Preferential Treatment	John Dovell Hartman	
6	Marginal Choices	Catharina G. M. Groothuis-Oudshoorn*	Juan Marcos González*, Dave Gebben, Marco Boeri
7	Super-stochastic Fantastic	Elisabeth Huynh	Akshay Vij, Habtamu T. Kassahun, Ali Ardeshiri, Flavio F. Souza, Subodh K. Dubey
8	Pio Pio	Juan-Manuel Ramos-Goni**	Oliver Rivero-Arias**, Mark Oppe*,**

* IAHPR Faculty members,
** Member of the EuroQol Group

Team Logos





Discreetly · Charming · Econometricians

Discreetly Charming Econometricians

Team Leader: Michał Kosma Jakubczyk

Team Members: Bogumił Kamiński, Dominik Golicki, Michał Lewandowski, Beata Koń, Paweł Ekk-Cierniakowski

Model Description:

1. Describe your choice of software and the reasons underlying your choice (e.g., Stata).

We used R. It's free and, hence, all the team members had access to it. We used custom-made estimation based on standard optimization techniques (we selected parameters' values to minimize a given goal function) and required no highly specialized packages (e.g., offering advanced panel data estimation techniques) that are not implemented in the R package ecosystem.

2. Describe your choice of estimation technique and the reasons underlying your choice (e.g., Bayesian).

We decided to use a very simple and straightforward approach. The actual reasons for the choices are most likely very complicated, full of heuristics, and not available even to the individuals making these choices. Therefore, we did not strive to really understand the choice mechanism, but only to get a reasonable fit within a model of a simple functional form. First, we neglected the fact that individual choices were available in the exploratory data. As we ultimately we need only to predict the rate of selecting a given profile in a given pair, we simplified the exploratory data to a single row per scenario (scenario = combination of profiles + left/right ordering, a profile = health state + duration). In addition, we found the statistical evidence for the effect of left-right ordering unconvincing and neglected it. In the estimation process, we used BFGS optimization algorithm with multi-start to find model parameters minimizing a given goal function in the exploratory dataset (we directly used the χ^2 goal function used by the organizers, in spite of some caveats—see the last point), as it seemed to offer most flexibility (not having to bother about linearity, etc.).

3. Describe your choice of functional form and the reasons underlying your choice (e.g., Logit).

First, we felt that choices and immediate death should be handled separately, and we built a linear model (RESET test suggested no misspecification). Regarding other scenarios, on the basis of past experience we calculated the relative attractiveness of available profiles and transformed it into rates (also to confine it to the [0,1] interval) with the arctan function (confirmed by the model fit, as compared to, e.g., tanh or logit). We allowed some parameters to vary with time unit based on the obvious fact that, e.g., absolute gains in years should matter more than the same absolute gains in days, but also that different time units might frame the problem differently. Statistical analysis and playing with various specifications suggested to us that, e.g., the importance of some dimensions may change with the time unit (e.g., due to usual activities being more important in the long-time horizon). At the same time, we tried to keep the number of parameters reasonably low, so we assumed that the relative weight of the levels is equal across dimensions. Finally, we decided that the actually observed frequency in a given scenario is the best prediction, so we combined the model predictions with these empirical rates whenever possible (with some additional minor tweaks).

4. Describe your choice of variables and the reasons underlying your choice (e.g., 20 effects-coded variables).

We assumed that respondents might find it difficult to compare options in terms of utility accumulated over time (as in standard QALY model), but might simply think of dimensions and duration as various criteria or attributes that have to be traded off. At the same time, we did not want to neglect thinking along the lines of the QALY model altogether. Therefore, we used three blocks of variables: (1) dimensions/levels of compared profiles (ignoring duration); (2) just duration of two profiles (compared in both an absolute and a relative way: i.e., 10 years vs. 8 years is both a 1.25 increase and a gain of 2 years. This was to handle, e.g., the possible non-proportionality); and (3) regular QALY component (utility-loss accumulated over time). As mentioned above, to keep the number of parameters low we assumed that the perception of levels (slight, moderate, etc.) is practically identical across dimensions (and then only multiplied by the importance of the dimension).

Modeling Recommendations:

5. Did you have difficulty modeling the 2 pair types (TTO pairs [quantity vs. quality] and efficient pairs [all attributes])? Did you have difficulty with the 4 temporal units (days, weeks, months, years)?

Frankly, we did not bother about TTO vs. efficient pairs differentiation in the modelling, hence: no. We feel that what matters more is whether two actual life profiles are being compared or a profile is being compared with immediate death.

We decided to allow some parameters to change with the time units used, as mentioned above. We had to decide which parameters should be allowed to change (to reflect the change in perception when talking about days or years) and which should be kept constant (to reduce the overall number of parameters). We were mostly guided by how the parameters changed when we were building separate models for each time unit: i.e., whether they changed in a monotonic way (suggesting some stable impact of time units) or behaved more haphazardly (suggesting noise only).

6. Do you believe that you would have been able to predict choice probabilities better had you received data on the respondent characteristics as part of the exploratory dataset (e.g., age)? Why?

No (assuming these characteristics are still not available in the prediction file, as the question seems to suggest). If these characteristics were available in the prediction file, that would require a different type of the prediction file in the first place (one row per single choice task, not per scenario), and that means an altogether different task.

7. Did you change your model's functional form or variables based on the estimation results (i.e., data mining)? If so, why and how? If not, why not?

Yes. We experimented with various specifications (e.g., independent estimation per time unit, relative or absolute impact of duration) and decided on whether goal-function improvement was big enough to increase the number of model parameters (we used the cross-validation to verify how big a decrease in the evaluation function could be treated as an actual improvement rather than simply the effect of a chance). We decided our function form was intuitive and simple enough for us not to run into the problem of over-fitting.

8. If your model wins, why do you believe it predicted better than the other models? If your model loses, why do you believe it did not predict better than the other models?

Case 1, we win. It is most likely (and this can be verified) caused, first, by our not being overconfident and using the actually observed frequencies whenever identical scenarios were used in the prediction set and exploratory data. Second, our noticing the impact of the time unit on the parameters (also the importance of some dimensions) might have helped us get a better fit. Third, questions of choices and of immediate death are qualitatively different, and we handled this in our approach. Fourth, we did not rely on the standard QALY model alone.

Case 2, we lose. First, as we wanted as much flexibility as possible, we did not use typical econometric software packages, so it was more difficult for us to conduct a standard statistical verification of our final model. That's why we went astray playing with various specifications based on intuition and only partially on statistical reasoning. Second, we might have been too modest to use not the model predictions but the observed frequencies for the scenarios that were in both sets (prediction and exploratory). The sample sizes (about 50 respondents) were too small, and model predictions are better, as they are based on much more data (and also on similar scenarios).

9. Based on your expertise and experience, what are the primary econometric advances needed to improve predictive modeling (not design)?

Well, the path we have taken clearly suggests that we do not really exploit the econometric toolbox intensively to predict choice. Therefore, we do not see any crucial advances. The design of the experiments is so much more important for grasping the factors (quantitative or qualitative) that affect the perception of the decision task and the attractiveness of the options at hand.

Competition Recommendations:

10. What recommendations do you have to improve the competition?

(1) Use more combination of health states times durations, and of time units times durations. Consider mixing various time units in a single scenario (e.g., months and years).

(2) Use a different evaluation function. The current one returns no value when $y_k = 0$ or $y_k = 1$ (using p_k in the denominator will not help, as a team might set $p_k = 0$ or $p_k = 1$). In addition, the current function might not promote truthful revelation: i.e., even knowing the true probability, it may still be optimal to report a modified value (to exploit the way the evaluation function changes with the actually observed frequency). We did not decide to take this path, however. (You may want to check Gneiting & Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," J Am Stat Assoc, 2007).

3) Participation-based financial reward is nice, but some element of competition would be even better.

4) That's an enormous undertaking, and so mistakes, etc., are unavoidable (great job, though!). Hence, more debugging is always needed (e.g., p_id 1129 and 1130 are identical, and so are 1133/1134 and 1154/1155. That might be of some importance when calculating the evaluation function—will you sum these scenarios twice or once?).

Occam's Barbershop Quartet

Team Leader: Kim Rand-Hendriksen

Team Members: Mathias Barra, Liv Ariane Augestad, Fredrik Dahl

Model Description:

1. Describe your choice of software and the reasons underlying your choice (e.g., Stata).

We prefer the statistical package R because it allows a full range of programming options and has support for an arbitrary number of data structures and functions, including structures such as multi-dimensional arrays. The fact that it is open-source and thus free is an added advantage.

2. Describe your choice of estimation technique and the reasons underlying your choice (e.g., Bayesian).

Model selection was based on leave-out cross-validation tests. Several basic approaches were considered for modeling the data, and we ended up creating a customized fitting function to allow testing of various non-linear models in the same framework and thus simplify cross-validation.

The chosen model has a certain resemblance to the Fedora example: it employs the Bradley-Terry model and uses a power function for time: Value = $\text{lifespan}^{\text{alpha}}$ - problems × $\text{duration}^{\text{beta}}$. However, the implementation differs, in that the functional form for problems is non-linear, with shared parameters for certain dimensions and levels.

3. Describe your choice of functional form and the reasons underlying your choice (e.g., Logit).

The Fedora way of handling time was elegant. We tested using separate models for the various time frames, but while this improved immediate model fit, cross-validation revealed that it reduced predictive accuracy. We tested various simplified, non-linear models (8-, 9-, 11-parameter), but found that these were under-fitted.

4. Describe your choice of variables and the reasons underlying your choice (e.g., 20 effects-coded variables).

If lowercase words represent dummy variables (mo2 = mobility at least level 2), the parameters, in addition to alpha and beta, are L3 (shared for mo, sc, pd, and ad), M4, M5, S4, S5, U3, UP5 (shared between ua and pd), U5, A4, A5, and DIFF45. DIFF45 is a parameter for number of dimensions at level 4 or 5.

MO * ((mo2 * (1-L3) + mo3 * L3) * (1-M4-M5) + mo4 * M4 + mo5 * M5) +

SC * ((sc2 * (1-L3) + sc3 * L3) * (1-S4-S5) + sc4 * S4 + sc5 * S5) +

UA * (ua2 * (1-U3-U4-UP5) + ua3 * U3 + ua4 * U4 + ua5 * UP5) +

PD * ((pd2 * (1-L3) + pd3 * L3) * (1-P4-UP5) + pd4 * P4 + pd5 * UP5) +

AD * ((ad2 * (1-L3) + ad3 * L3) * (1-A4-A5) + ad4 * A4 + ad5 * A5) + s45 * DIFF45

Modeling Recommendations:

5. Did you have difficulty modeling the 2 pair types (TTO pairs [quantity vs. quality] and efficient pairs [all attributes])? Did you have difficulty with the 4 temporal units (days, weeks, months, years)?

We are used to modeling continuous data, so DCE pairs were a challenge. The use of various temporal units complicated things further. It took more time than anticipated to create a framework for performing cross-validation on the models we wanted to test.

6. Do you believe that you would have been able to predict choice probabilities better had you received data on the respondent characteristics as part of the exploratory dataset (e.g., age)? Why?

Possibly, but probably not by much. We did consider creating a function that would allow fitting of separate functions for different sub-groups of the sample population, but the approach we had in mind would have been possible without access to respondent characteristics. While it may be interesting to figure out what kind of variation there is in response style depending on respondent characteristics, we do not see this as crucial for generating value sets for instruments such as the EQ-5D, since we wish to employ the same value set for all.

7. Did you change your model's functional form or variables based on the estimation results (i.e., data mining)? If so, why and how? If not, why not?

Cross-validation can be seen as a form of data mining. While cross-validation is intended to reduce the risk of over-fitting, it is still possible to do so given a sufficiently large sample of candidate models.

8. If your model wins, why do you believe it predicted better than the other models? If your model loses, why do you believe it did not predict better than the other models?

The model is probably simpler, in that it employs relatively few parameters. The addition of the DIFF45 term improved predictive ability substantially. We have not looked closely at other possible interaction terms, so there is a chance that we have overlooked another term that could have improved the model further.

9. Based on your expertise and experience, what are the primary econometric advances needed to improve predictive modeling (not design)?

Non-linear models are still relatively difficult to implement and can be computationally challenging even with modern computers. Improvements in general-purpose, unrestricted modeling tools could result in substantial improvements.

Competition Recommendations:

10. What recommendations do you have to improve the competition?

The competition would have been even more interesting if it involved more than two consecutive samples, giving the participants the option of altering their models and choices as more information was made available. The addition of respondent-characteristic information could result in interest from a wider audience.

Basta!

Team Leader: Mathias Barra

Team Member: Liv-Ariane Augestad

Model Description:

1. Describe your choice of software and the reasons underlying your choice (e.g., Stata).

Team Basta has used R (R Core Team 2015). It is free, and you can do anything you like.

2. Describe your choice of estimation technique and the reasons underlying your choice (e.g., Bayesian).

We kept it simple. As we believe there is no a priori reason to expect any particular framework (e.g., RUT), we decided to try to model possible heuristics respondents might use. Assuming a 50-50 point of origin, we iteratively tried to identify possible heuristics. Manually inspecting random portions of the exploratory data, we tried out various specifications, and stepwise added or eliminated predictors and interactions. As an example, we hypothesized that there would be some differences between heuristics used for pairs with only two mild states (all dimensions at level 3 or below). A dummy flagging such pairs was computed, and it interacted with a variable encoding the difference in duration of the two alternatives.

3. Describe your choice of functional form and the reasons underlying your choice (e.g., Logit).

Not applicable.

4. Describe your choice of variables and the reasons underlying your choice (e.g., 20 effects-coded variables).

Our model is of the form p - 0.5 ~ X with standard OLS-modeling. The offsetting of the observed proportions by –0.5 is to facilitate the interpretation of the coefficients and to enforce an intercept-less model. The result is that a 50-50 distribution—i.e., p = 0.5—is represented as 0, a negative value means that more respondents chose B, and a positive value means that more respondents chose A. Several of the predictor-variables were therefore signed. For example, one of the retained predictors was a variable DELTA_T encoding the absolute difference in duration between the two alternatives (viz., T_A – T_B). This predictor is positive [negative] when the duration of alternative A is longer [shorter] than the duration of alternative B. Hence, when the coefficient for this predictor is alpha > 0, then if, e.g., the duration of B is 2 units of time longer than the duration of A, our model predicts that the observed proportion should be p – 0.5 + (–2) * alpha. In this case DELTA_T is –2, which means that the prediction is adjusted downwards, as explained above, in favor of B. After several iterations of the process of trying out new predictors, inspecting pairs with large residuals for discernible patterns, and adding new ones that could explain these, we reached our preferred specification with the following predictors: (1) a predictor for severity-difference, which was computed as a the difference of weighted sums of dummies for each dimension at or above 4 for the two alternatives; (2) a secondary-misery-difference of a weighted sum of dimensions at or below 3; (3) a variable time-difference (DELTA_T, described above); (4) a relative-time-gain predictor equal to the signed scalar sign (DELTA_T) * max(T_A,T_B)/min(T_A,T_B); (5) a signed perfect-health dummy, coding pairs in which A [B] was 11111 as 1 [–1]; (6) a milds-dummy (described above); (7) a signed dominance-dummy set to 1 [–1] when A [B] was dominating, in the sense that the severity-score (predictor (1)) was lower *and* relative-time-gain was not at or above 1.5 for B [A]; and finally (8) an extra severity-difference predictor fives-difference computed as a weighted sum of dummies for dimensions at level 5. The milds-dummy (6) was used only

as an interaction term with the time-delta and the secondary-misery-difference variables so that the final specification can be described as (1) + (4) + (7) + (8) + (2):(6) + (3):(6) .

Modeling Recommendations:

5. Did you have difficulty modeling the 2 pair types (TTO pairs [quantity vs. quality] and efficient pairs [all attributes])? Did you have difficulty with the 4 temporal units (days, weeks, months, years)?

We treated neither type (TTO or efficient-pair) very differently. The TTO-pairs were to a certain extent accounted for by the perfect-health-dummy. However, the sign of this coefficient surprisingly (?) was negative, meaning that respondents (under our modeling assumptions) are more likely not to choose the 11111 alternative. This might be interpreted as representing non-traders (because it so clearly means giving up time.) With respect to the temporal modalities, we specified and estimated one model for each modality, albeit with the same predictors. However, the weights used to compute the two severity-difference predictors were calibrated separately for each modality. (This was done with an external optim-procedure after initial experimentation with guesstimated weights—space does not allow for a complete description of the calibration here.)

6. Do you believe that you would have been able to predict choice probabilities better had you received data on the respondent characteristics as part of the exploratory dataset (e.g., age)? Why?

Yes, slightly, depending on what kind of additional variables would be available. Most studies find that there are systematic reproducible and theoretically meaningful differences in choice-behavior and -preferences between subgroups.

7. Did you change your model's functional form or variables based on the estimation results (i.e., data mining)? If so, why and how? If not, why not?

Yes, as described above, quite a lot of tweaking and trying-out was done. Indeed, the whole concept of the model was manually inspecting pairs and trying to guess the observed proportions (actually the deviations from 0.5) and then implementing these as dummies or heuristically computed measures of differences between the two alternatives of a pair. In addition, in the course of selecting the weights for the severity-difference-variables, we initially hypothesized weights based on introspection and the literature on hypothetical evaluations (which worked quite well) before we invested some effort in optimizing these once a predictor had been chosen for inclusion. We initially included an intercept (which was almost 0 and thus also served as a sort of internal validation), but this was omitted before the final predictions because it is theoretically meaningless, as the pairs' alternatives are randomized to A and B prior to presentation to the respondents.

8. If your model wins, why do you believe it predicted better than the other models? If your model loses, why do you believe it did not predict better than the other models?

We are quite confident that our model will not finish last: It performs quite well for all internal validation and is nicely monotone and uniform between the temporal modalities. If it loses (we take this as not winning and as not finishing last), we will blame bad luck. If it wins, it is probably because it makes no appeal to any underlying utility-assumptions or other hypothetical constructs, and (in our opinion) tries to tackle the challenge head-on. We try not to predict the alternatives' values, but rather the inclination of the respondents to choose one or the other based on characteristics of the pair. Of course, this distinction is not clearly demarcated for the others, but the perfect-health dummy, the relative time scalar, or the domination-dummy should serve as examples with little appeal to other than basic choice-heuristics.

9. Based on your expertise and experience, what are the primary econometric advances needed to improve predictive modeling (not design)?

Back to basics! If one wants to have a ratio-scale utility for health benefits, one must understand that value is nothing but the value that individuals attach to the thing. If one wants to do CUA, a theoretically and empirically feasible way of representing individuals' preferences on a ratio scale is necessary to make real advances.

Competition Recommendations:

10. What recommendations do you have to improve the competition?

We believe that the outcome measure (the "chi-squared") is very arbitrary because it so strongly favors a model that accurately predicts the fringe probabilities (those close to 0 or 1). There is no reason these should be more important (unless one has decided up front that the probabilities will ultimately be employed to derive utilities—something that can be questioned both on theoretical and empirical grounds). Instead, a more neutral fit-metric (e.g., simply sum((pred – obs)^2)) could be used. In fact, this is the metric we used for cross-validation purposes, etc., during the model-selection, and we have even hedged our predictions against the chi-square measure by computing a transformation designed to minimize the expected chi-square measure for our predictions, taking advantage of the fact that if the residuals (in our model) are normally distributed around the predictions, then on average one gains more by shifting predictions slightly away from 0.5 toward the more extreme probabilities. In fact, the chi-square for the exploratory data was improved by approximately 16% with this hedging shift.

Fedora

Team Leader: Benjamin M. Craig

Team Members: None

Model Description:

1. Describe your choice of software and the reasons underlying your choice (e.g., Stata).

For this competition, I used the STATA MP, specifically version 14.1 with 12 cores. From my perspective, this software has provided me with a suitable balance between packaged commands and the ability to program and publish my own commands (e.g., hyreg). Knowing what the estimator is doing at each step in the analysis has provided me with greater insights into the limitations of alternative approaches (such as the importance of initial values). Although I read the descriptions provided in the software handbooks, I also appreciate seeing how the code operates (line by line), which is not possible with many canned packages. At times, the small numerical issues (such as rounding errors) can have substantial consequences for the interpretation. I am trained to program in other languages, but some require more effort for simple commands (e.g., Gauss) or don't show the underlying code (SAS). Other reasons for using this software package include its handling of large datasets, its accessibility for novice programmers, its widespread community of users, and its graphics.

2. Describe your choice of estimation technique and the reasons underlying your choice (e.g., Bayesian).

As a frequentist, I typically begin modeling by specifying a decision rule to either maximize or minimize. In the case of binomial analysis, I choose weighted least squares (WLS), because this has the same first derivative as the maximum likelihood (ML) function but allows for unanimous predictions (e.g., 0 and 1). The weights are typically based on the predicted probability, not the empirical probability (See response 5). WLS has the limitation that if the predictions are unanimous, a weight correction is required (e.g., Berkson weights); otherwise, it has served me well in past analyses. This estimation technique is also known as minimized chi-square, Urban's Normit, or GLM. Basically, it has the advantage of minimizing chi square, which is the basis of this competition. If it is the winner, I may explore alternative estimation techniques that allow preference heterogeneity, which will require switching to either ML or Bayesian approaches (i.e., more assumptions).

3. Describe your choice of functional form and the reasons underlying your choice (e.g., Logit).

The functional form has two components. The first is the cumulative density function (CDF). For this competition, I choose to use the Bradley-Terry model (i.e., A/(A+B)). Under this CDF, scaling terms that are common to A and B cancel. The second component is the value specification. For this, I choose Value = $\text{lifespan}^{\text{alpha}}$ - problems × $\text{duration}^{\text{beta}}$, where *problems* includes the 5 attributes of the EQ-5D description. This functional form was identified semi-parametrically in a previous study. In this competition (unlike the previous study), all problems have the same duration as lifespan (i.e., lifespan = duration), but it seems appropriate to allow for different time preferences (alpha and beta). If I were to choose an alternative functional form, I would consider allowing the beta to vary by health problem (i.e., the effect of the duration of slight problems may not be the same as for severe problems).

4. Describe your choice of variables and the reasons underlying your choice (e.g., 20 effects-coded variables).

The regression model includes only the 20 effects-coded dummy variables. These are standard in most EQ-5D valuation studies. In this parsimonious model, each coefficient represents the loss in QALYs incurred by an increase in a domain by 1 increment of level (e.g., going from level 2 (slight problems) to 3 (moderate problems) on mobility). I did not include any interaction terms, adjustments for scale, adjustments for temporal units, adjustments for pair types (i.e., TTO pair vs. efficient pairs), or behavioral parameters (e.g., left/right, sequence), which might have improved fit. This is the simplistic approach (20 regression parameters and 2 time-effect parameters [alpha and beta]).

Modeling Recommendations:

5. Did you have difficulty modeling the 2 pair types (TTO pairs [quantity vs. quality] and efficient pairs [all attributes])? Did you have difficulty with the 4 temporal units (days, weeks, months, years)?

No, I did not. Studies on preferences between health-related goods and services (e.g., choice-based conjoint) tend to be small and focused: for example, the value of a night in the hospital after knee surgery. In health valuation, we attempt to understand preferences on all health outcomes (e.g., imagine all possible durations and experiences in a hospital). Sometimes we ask about preferences between improved quality of life and extended lifespan (i.e., QALYs using TTO pairs) and other times we ask about preferences between two entirely different health outcomes (i.e., efficient pairs). Both are relevant, therefore we need a unifying model (e.g., Fedora).

6. Do you believe that you would have been able to predict choice probabilities better had you received data on the respondent characteristics as part of the exploratory dataset (e.g., age)? Why?

I do not believe so. Preference heterogeneity exists, but it is difficult to separate preference heterogeneity from blocking in the pair allocation (e.g., respondents asked similar pairs may seem to have similar preferences). Little evidence suggests that there are substantial differences in preference weights across HRQOL domains, but there might be differences in time preferences (alpha and beta) by respondent age and health.

Most estimation techniques applied to identify preference heterogeneity are assumption-laden and -dependent. A natural next step after identifying the merits of alternative modeling approaches would be to examine which models predict heterogeneity the best (similar to a confirmatory factor analysis). This falls outside the scope of the competition at this time.

7. Did you change your model's functional form or variables based on the estimation results (i.e., data mining)? If so, why and how? If not, why not?

Yes, I changed the functional form in two ways. First, I originally planned to include a theta term to adjust for non-traders in the analysis (i.e., the proportion of respondents who always choose the longer lifespan); however, the parameter added little to the predictions and was dropped. Second, I originally ran WLS using the empirical probabilities in the sample weights (see example code), but I replaced these with the predicted probabilities, because the revised estimation is more efficient when correct. This second change had little to do with the results, but was based on a discussion with Mark Oppe. The parameters and variables were not changed based on the results.

8. If your model wins, why do you believe it predicted better than the other models? If your model loses, why do you believe it did not predict better than the other models?

The top three reasons that Fedora might win are: (1) It does not use a logit, and therefore is not as susceptible to proportional scaling issues; (2) it relaxes the constant proportionality assumption by incorporating time preferences (alpha and beta); and (3) it minimizes chi square, which is the primary metric for model comparison.

The top three reasons that Fedora might lose are: (1) It does not data mine, and a model with more parameters may predict better (e.g., N3 term); (2) it does not take into account behavioral effects, such as left/right bias, sequence bias, and non-trading, which could reduce error and improve prediction; and (3) it does not separately model the TTO pairs. The TTO pairs have the same number of attributes but require less information because one description is always in full health. Modeling these choices may be intrinsically different from modeling the more complex paired comparisons. As discussed here, Fedora is limited to just preference attributes and does not incorporate behavioral characteristics, which might predict choice beyond preference.

9. Based on your expertise and experience, what are the primary econometric advances needed to improve predictive modeling (not design)?

The top three advances to improve predictive modeling are: (1) We need a better understand of the correlation between choices (i.e., two choices from one person might provide more information than two choices from two people [e.g., intervals]); (2) we need greater diagnostic tests and investigation in interaction effects; and (3) we need to examine angle- and ratio-based approaches to modeling choice (e.g., log Cauchy, Bradley-Terry), which may address censoring and scaling issues.

Competition Recommendations:

10. What recommendations do you have to improve the competition?

I wish that someone else would run the competition and that I did not have to share my entry in advance (joke). This is our first time running such a competition, and Kim and I are quite pleased with how it is going so far. We are particularly grateful for the wealth of support from the teams. Nevertheless, I would recommend for next time that (1) we have multiple rounds, kind of like playoffs; (2) we improve the design based on the victorious model (20/20 hindsight); and (3) we have more prizes (e.g., most elegant model), maybe even a prize for the worst model (i.e., Lanterne Rouge). Mostly, I want us to have geeky fun.

Preferential Treatment Submission

Team Leader: John Hartman

Team Members: None

Model Description:

1. Describe your choice of software and the reasons underlying your choice (e.g., Stata).

I used STATA 13.1 for my submission. I chose STATA as it is the package I am most familiar with and found the example code easy to modify to fit my needs.

2. Describe your choice of estimation technique and the reasons underlying your choice (e.g., Bayesian).

I chose to use weighted least squares, as I found that other methods resulted in a larger chi-square and were thus inferior for the purposes of this competition.

3. Describe your choice of functional form and the reasons underlying your choice (e.g., Logit).

As a student of Dr. Craig, I have become convinced that the Bradley-Terry model offers some advantages over a logit model. From the results of a previous study as well as numerous conversations with Dr. Craig, I also believe that there is an alpha and beta function for the valuation of health (value = $\text{lifespan}^{\text{alpha}}$ – problems * $\text{duration}^{\text{beta}}$). For my model, I specified both alpha and beta equal to .45.

4. Describe your choice of variables and the reasons underlying your choice (e.g., 20 effects-coded variables).

My regression model includes both the 20 effects-coded dummy variables and two additional variables that I think may affect choice. The first is a sort of "pit state" dummy equal to 1 if the sum of the five domain levels of the EQ-5D-5L is >17. The second is another dummy variable equal to 1 if one of the two choices has a lower sum of total problems and has a longer lifespan. I believe that these are two methods that individuals use to "judge" differences between choices. In the future, I would like to spend more time looking at some of the heuristic patterns that individuals follow in their choices.

Modeling Recommendations:

5. Did you have difficulty modeling the 2 pair types (TTO pairs [quantity vs. quality] and efficient pairs [all attributes])? Did you have difficulty with the 4 temporal units (days, weeks, months, years)?

One problem I found was that the TTO pairs didn't cover the same levels as the efficient pairs. I believe that six of the 25 total domain levels (including level 5 for most domains) were not included in the TTO pairs. I believe that the TTO pairs may be easier for respondents to understand and may be more precise, especially when one of the options has severe or extreme problems. I had no difficulties with the 4 temporal units.

6. Do you believe that you would have been able to predict choice probabilities better had you received data on the respondent characteristics as part of the exploratory dataset (e.g., age)? Why?

I think the inclusion of respondent characteristics would provide minimal improvements to my predictions. I am currently in the process of submitting a paper comparing differences in valuation between parents and non-parents and have found significant differences between the two. I think that

including variables on individual income, education, and parental status might provide slight improvements to my model.

7. Did you change your model's functional form or variables based on the estimation results (i.e., data mining)? If so, why and how? If not, why not?

Yes. I originally ran my model using only the 20 parameters for differences in health states along with the alpha and beta terms. I then included two additional parameters (pit states and better states with a longer lifespan) and experimented with others, among them separating the TTO and efficient pairs and including a time dummy when there was a large difference in lifespan between the two health states. I then compared the model fit for each set of parameters to decide which one would be chosen for my submission.

8. If your model wins, why do you believe it predicted better than the other models? If your model loses, why do you believe it did not predict better than the other models?

While I believe that it is highly unlikely that my model will win, I believe that it may not come in last place because I relaxed the constant proportionality of the time assumption and chose a modeling method (Bradley-Terry) that is not as common as the logit. Possible reasons that my model may lose are the fact that I fixed the alpha and beta terms to a specific value, that I did not include many additional parameters, and that I didn't control for any individual response patterns.

9. Based on your expertise and experience, what are the primary econometric advances needed to improve predictive modeling (not design)?

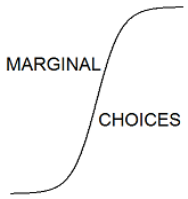
As a junior researcher and having the least modeling experience, I believe that competitions such as this one will point me in the right direction for future research. If one model is found to dominate the others, then this can be used as a base to explore how and why it performed better than the others.

Competition Recommendations:

10. What recommendations do you have to improve the competition?

Overall, I think that this whole process has been a great experience. I greatly appreciate the example code that was provided and know that it has greatly benefitted my understanding of choice modeling and saved me quite a few headaches and hours of frustration. I think that allowing individuals who didn't even enter the competition to have access to the sample code is also a great idea.

Marginal Choices



Team Leader: Karin Groothuis-Oudshoorn

Team Members: Juan Marcos González, David Gebben, Marco Boeri

Model Description:

1. Describe your choice of software and the reasons underlying your choice (e.g., Stata).

For this competition, we used primarily STATA MP 14.1 with 4 cores. Our team chose STATA partly because all the members of the group were familiar with the software, but also because our model was an adaption of the Fedora model in the sample codes included in the competition materials. The team also used a kind of cross-validation of the exploratory data to contrast a number of different model specifications. Because we had used STATA for the modelling, the simulations were also done in STATA.

2. Describe your choice of estimation technique and the reasons underlying your choice (e.g., Bayesian).

The most optimal solution of the contest will be based on comparing predicted and observed probabilities with a chi-square for each comparison of two health states. These probabilities are aggregated over respondents, and therefore we did not use an individual model—that is, a model that accounts for heterogeneity between respondents. For the same reason, we think that a Bayesian technique would not be more advantageous because no respondent characteristics were available.

A basic model would be a conditional logit with a specification of the utility of a choice set. However, as was shown by the Fedora team, this model did not perform very well.

The Bradley-Terry model, based on minimizing the chi-square of comparing the estimated with the observed probabilities using a weighted least-squares estimation performs better. In Berkson (1950), it was shown that for estimating a proportion, the error variance of the estimator of a weighted least-squares estimation is lower than that of the maximum-likelihood estimator.

The dependency of the estimated parameters on the estimation procedure for the maximum-likelihood estimation was checked. Other available algorithms in STATA gave the same parameter estimates and log likelihood value, but the convergence was quickest for the NR algorithm.

3. Describe your choice of functional form and the reasons underlying your choice (e.g., Logit).

The basis of our model was the value specification of the Fedora team:

$$\text{Value} = \text{lifespan}^{\text{alpha}} - \text{problems} * \text{duration}^{\text{beta}}$$

and the Bradley-Terry model.

We checked different specifications for alpha and beta as function of characteristics of the choice sets. Our final model is:

$$V = \text{lifespan}^{\alpha_0+\alpha_1 \text{TTO}+\alpha_2 \text{time}_1+\alpha_3 \text{time}_2+\alpha_4 \text{time}_3} - \text{problems} * \text{duration}^{\beta_0+\beta_1 \text{TTO}+\beta_2 \text{time}_1+\beta_3 \text{time}_2+\beta_4 \text{time}_3},$$

where TTO = 1 if the choice set was a TTO pair, and 0 if the choice set was an efficient pair. “Duration” is a linear function of losses in utility induced by an increase in a health domain by an increment from the level. So each attribute with 5 levels of the EQ5D is represented with 4 dummy variables. The final model contains 20 + 10 = 30 parameters.

If we were to choose an alternative model, we would consider the Thurstone-Mosteller model, which is an alternative to the Bradley-Terry model based on a normal distribution. But usually the two models lead to similar estimates.

4. Describe your choice of variables and the reasons underlying your choice (e.g., 20 effects-coded variables).

We investigated several different model specifications based on the information available in both the exploratory and confirmatory data. We identified the following observable pieces of information: Severity of quality-of-life outcomes, duration of quality-of-life outcomes, and question type (i.e., TTO versus efficient pairs).

The basis for the model specification used was the model proposed by the Fedora team. Two types of extensions were considered to relax some implicit assumptions in the model proposed by the Fedora team: extension of the specification of the problems = health state, and extensions of the model for beta/alpha. Although we studied several extensions, we following are mentioned:

A. Including a severity variable in the specification of the "problems": a dummy variable indicating whether the profile contained a level larger than 3, or a dummy variable indicating whether the profile contained a level larger than 4. This did not lead to a better log likelihood (severity larger than 3: -2532.4; larger than 4: -2533.0).
B. Including an additional constant effect adjusting choice probabilities by pair type, TTO pair vs. efficient pair.
C. Including a categorical variable for temporal units (time_1 = days, time_2 = weeks, time_3 = months, time_4 = years). In addition to affecting the timing of the tradeoffs evaluated in each question (i.e., at least one year in the future, at least one month in the future, etc.), which are considered in the discounting effect in the original Fedora model, the temporal units also determine the minimum amount of time traded, which potentially influences the closeness of the pairs evaluated in terms of well-being. It also potentially systematically affects choice probabilities. In the case when one profile contained "immediate death," the temporal unit of the other profile was taken, because "immediate death" occurred in only 40 choice sets, and including an extra level could then lead to a badly specified model due to sparseness.
D. Combination of B and C.

Finally, we focused on the models in B (24 parameters), and D (30 parameters). We will call these models from now on mc1 and mc2. The model we used to predict the proportions was mc2.

Modeling Recommendations:

5. Did you have difficulty modeling the 2 pair types (TTO pairs [quantity vs. quality] and efficient pairs [all attributes])? Did you have difficulty with the 4 temporal units (days, weeks, months, years)?

First, we wanted to check whether the two types of pairs could be combined in one dataset. However, when modelling the two types of pairs separately we could not fit a Bradley-Terry model for the TTO pairs, and no convergence was obtained. We could separately fit a conditional logit model on the TTO pairs and the efficient pairs. It turned out for the log likelihoods that $LL_{\text{efficient pairs}} = 40557.2$, $LL_{TTO} = 13746.7$ and $LL_{total} = 54353.2$. Because the total sum of the loglikelihoods of the separate datasets (= 54303.9) was not greatly different from LL_{total}, we concluded that there seems to be no problem with analyzing the data simultaneously. Moreover, we estimated a relative scale parameter between the TTO pairs and the efficient pairs for (1) the constant proportionality model, (2) the Fedora model, and (3) our

model. It turns out that introducing the scale parameter doesn't cause the chi-square to change much (2528.9, as compared to 2533.21, the chi-square for our model), and it turns out that the likelihood as a function of the scale is quite flat. The optimal value for the scale parameter is 1.21.

6. Do you believe that you would have been able to predict choice probabilities better had you received data on the respondent characteristics as part of the exploratory dataset (e.g., age)? Why?

We do think that additional personal and attitudinal information could be important ingredients for a better prediction of choice probabilities. To the extent that observable characteristics are related to preference heterogeneity, the accuracy of out-of-sample predictions can be partly related to the representativeness of the study sample. Future draws on the population can be systematically different in the mix of preferences elicited, which in turn could make average estimated preferences—and predicted choices—from previous samples poor prediction tools. Although large sample sizes can minimize sampling issues, determining whether the size of the current samples—both exploratory and confirmatory—are enough to approximate the preferences of the population is ultimately an empirical question. Our team used splitting of the exploratory sample to simulate the effect that deviations in unobserved characteristics could have on the accuracy of out-of-sample predictions. We found quite a large dispersion in prediction accuracy with several model specifications, suggesting that auxiliary information could indeed have improved our ability to predict choice probabilities.

7. Did you change your model's functional form or variables based on the estimation results (i.e., data mining)? If so, why and how? If not, why not?

The chi-squared value and the proportions of rejected predicted probabilities of the model mc2 were smaller than these outcomes for model mc1. And the log likelihood difference between those two models was significant (namely, 2533.2 - 2505.1 = 28.1 with 6 df).

We checked the predictive behavior of these two models based on splitting the sample of respondents. We divided the sample in half, with equal numbers of the different time slots in both subsamples, and estimated model mc1 and mc2 on one half of the data, the inner sample. Then we calculated the chi-square value and the number of rejected points for the other half of the data (the outer sample). We repeated this 1000 times.

It turns out that the number of rejected points is smaller for mc2 in 429 cases, but smaller for mc1 in 408 cases (and equal for 163 samples). The chi-square value is smaller for 387 of the samples for model mc2. So it could be that model mc2 is only marginally better than model mc1.

The 95% confidence interval of chi-square for mc1, based on empirical distribution, is 2861.7 - 3589.2. For model mc2 it is 2922.0 - 3581.4.

Histograms of the outcomes can be found in the corresponding zipfile (boot_chisquare_mc1.png, boot_reject_outsample_mc1.png, boot_chisquare_mc2.png, boot_reject_outsample_mc2.png).

8. If your model wins, why do you believe it predicted better than the other models? If your model loses, why do you believe it did not predict better than the other models?

If we lose it could be due to chance, since in part the ranking will be based on the validation dataset, which is only one dataset. As we saw from the split sample simulation, the variability of the performance of the models between datasets is quite large. Although we were able to make some improvement over the model of the Fedora team, it was, in our opinion, only a moderate improvement given the limited number of additional pieces of information provided. Thus if a better model exists, we think it would be based on a completely different specification of the value function or the underlying estimator. We have

searched for one within the time constraints of the competition, but we couldn't find a better alternative than the Bradley-Terry model.

9. Based on your expertise and experience, what are the primary econometric advances needed to improve predictive modeling (not design)?

Explore non-parametric models for predictive modeling: for example, based on nearest-neighbor methods. In nearest-neighbor methods, one has to define a distance measure for the choice sets and then predict the probability for a new choice set on the observed probabilities of neighbor choice sets. Choosing a distance measure is similar to specifying the value function, but the class of distance measures is much broader.

Competition Recommendations:

10. What recommendations do you have to improve the competition?

First of all, we would like to say that this competition was and is great fun, and we would like to thank the organizers of the competition. It gave us a chance to work on an international team, and we learned a lot about searching outside our comfort zones for different models.

We should look for a criterion for deciding which model wins or loses independently of observed data. It could be that a certain model outperforms the others, but for another dataset the ranking could be different due to chance.

References: Berkson J. Relative precision of minimum chi-square and maximum likelihood estimates of regression coefficients. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. July 31-August 12, 1950. Statistical Laboratory of the University of California, Berkeley. Berkeley, Calif.: University of California Press, 1951. 666 pp. Editor: Jerzy Neyman, p.471-479

Super Stochastic Fantastic

Team Leader: Elisabeth Huynh

Team Members: Akshay Vij, Habtamu T. Kassahun, Ali Ardeshiri, Flavio F. Souza, Subodh K. Dubey

Model Description:

1. Describe your choice of software and the reasons underlying your choice.

Several different softwares were utilized for analysis, these ranged from standard statistical packages such as STATA to fully developed programming languages such Python. The final model submitted for this competition was estimated in STATA MP 13. This software is a complete, integrated package that is user friendly as is one of the most popular commercial packages in the handling of large datasets, for analysis, and its graphics. It has a great degree of flexibility, seemingly allowing the user to transition between a GUI interface, packaged commands and the ability to program and estimate their own models. Stata offers a wide range of statistical analyses that range from estimating simple linear regression models to more generalized methods of moments estimations. The estimation of packaged commands are fast and efficient and is well controlled by StataCorp so that the results are reliable. Moreover, there is a widespread community of users that share their open source codes and own packages that can be directly installed for use.

2. Describe your choice of estimation technique and the reasons underlying your choice.

We explored a number of different classification models popular in the machine learning literature, such as decision trees, neural networks, support vector machines, etc. However, we quickly realized that the strength of these information theory based classifiers lies with problems where the vector of explanatory variables is of the order of hundreds, if not thousands, compared to six variables in our case. Therefore, we made the decision to use a simpler model and expend more energy in identifying the most appropriate model specification.

To account for the proportional response dependent variable (aggregate choice shares for choosing alternative 1 over alternative 2 for a particular pair), we estimated a number of censored regression models, transformed models and fractional response models that account for the natural censoring of the data at zero and one. The final model is a fractional logit model (Papke and Wooldridge, 1996 Journal of Applied Econometrics) which is estimated as a generalized linear model with a binomial family and logit link function specification. Papke and Wooldridge (1995) show that maximum likelihood standard errors are too large, the fractional logit model can be estimated quite straightforwardly via quasi-maximum likelihood estimation using the glm command in STATA and assuming robust standard errors.

3. Describe your choice of functional form and the reasons underlying your choice.

For reasons mentioned in response to the previous question, we decided to utilize models that account for the proportional response dependent variable. Furthermore, we didn't see much benefit in exploring different functional forms for the stochastic component of the utility specification (we saw greater value in exploring the appropriate specification for the systematic component). Therefore, we chose the simplest functional form of these models.

4. Describe your choice of variables and the reasons underlying your choice.

We conducted some preliminary analysis through cross-tabs and such to understand if there was some structure to how individuals were choosing between scenarios. We noticed that the processing of attributes related to mobility, self care, activity, pain and anxiety seemed to depend on how much time the individual was told they had to live. This made immediate sense. One's choices when one expects to live several years will likely differ from one's choices when one expects to live only a few days or weeks. The patterns that we saw seemed to indicate that different modes of decision-making may be activated, depending on whether an individual expects to live a few days, weeks, months or years. Therefore, we segmented the data based on whether expected lifespan was presented in units of days, weeks, months or years, and we estimated four separate models on each of these four subsets.

For the sake of consistency, we maintained the same specification across all four segments. Each of the five quality-of-life variables: mobility, self care, activity, pain and anxiety, were dummy-coded and included in the utility specification. Following Viney et al. (2014 Health Economics), we also explored non-linear preferences with respect to time and interactions between the EQ-5D attributes with time. Time was specified as having a quadratic effect and interactions of time with continuous levels of the quality-of-life variables were also included in the specification. To account for potential cross-effects (that is the effect of the other alternative on choosing the current alteranative), we also included cross-effects for all the variables described above (five quality of life variables, quadratic time and time interaction variables).

Finally, we also included a binary variable for cases where a scenario results in immediate death, since we expected such scenarios to have a greater disutility than would otherwise be captured by a specification that is merely linear and quadratic in expected lifespan.

To summarize, we estimated 56 parameters per model, and we estimated 4 models in total, one each for when expected lifespan was presented in units of days, weeks, months and years, resulting in a total of 224 parameters.

Modeling Recommendations:

5. Did you have difficulty modeling the 2 pair types (TTO pairs [quantity vs. quality] and efficient pairs [all attributes])? Did you have difficulty with the 4 temporal units (days, weeks, months, years)?

With regards to the TTO pairs, we did explore the idea of modelling them separately, but we did not find much benefit from adopting such an approach. As mentioned earlier, we also explored the idea of segmenting the data by temporal units, and we found it to yield much better results.

6. Do you believe that you would have been able to predict choice probabilities better had you received data on the respondent characteristics as part of the exploratory dataset (e.g., age)? Why?

Absolutely! Individual preferences have been found to vary systematically as a function of demographic characteristics across a wide variety of empirical contexts, and there do not seem to be strong reasons to believe the same wouldn't be true in this particular context. Somebody who's 50 years old may evaluate expected lifespans very differently from someone who's 70 years old. Mobility might be more important to individuals who have led relatively active lives, which in turn might be correlated with observable demographic variables, such as occupation (sedentary vs. active professions).

7. Did you change your model's functional form or variables based on the estimation results (i.e., data mining)? If so, why and how? If not, why not?

Yes, multiple times. Since the goal was to maximize prediction accuracy over 3200 choice scenarios, of which only 1560 were included in the estimation (or training) data, we made extensive use of validation and cross-validation techniques, where a subset of pair IDs was not used for estimation.

We started out by exploring quite complex models, with parametric and nonparametric univariate and multivariate mixture distributions. While these models fit the data much better, we found improvements in prediction accuracy modest at best.

This led us to revise our approach and focus more on the specification of the systematic component. We tried a number of specifications where different subsets of the variables were interacted, but again, we did not see a marked improvement in prediction accuracy.

In the end, segmenting by units of time and allowing for cross-effects worked the best in terms of prediction accuracy. And from a behavioral standpoint, the specification made a lot of sense, since it is quite likely that different modes of decision-making might be activated, depending on whether an individual expects to live a few days, weeks, months or years, and cross-effects would affect choices.

8. If your model wins, why do you believe it predicted better than the other models? If your model loses, why do you believe it did not predict better than the other models?

We believe segmenting the data by units of time and allowing for cross-effects was perhaps our most original, and retrospectively obvious, idea, and if we win, it would be on the strength of this idea. However, we relied on the idea at the expense of more complex structures for the stochastic component, and if we lose, it would probably be because we overlooked particular error structures.

9. Based on your expertise and experience, what are the primary econometric advances needed to improve predictive modeling (not design)?

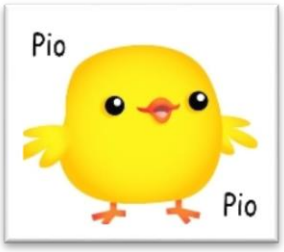
We believe advances in machine learning and data sciences hold important lessons for econometricians and statisticians who work in predictive modelling. However, the disciplines of machine learning and data sciences often focus much more on prediction and not as much on model inference and interpretation. While this might be an acceptable practice under certain contexts, where for example the prediction data is not very different from the training data, in cases where the prediction data is expected to be markedly different, such as when predicting the impact of new technologies or services, or forecasting over long-term horizons of the order of decades, better tools for model inference and interpretation could help guide the process of model development in ways that just wouldn't be possible using prediction metrics alone. Therefore, we believe that greater cross-fertilization between statistics and econometrics on one hand, and machine learning and data sciences on the other, could greatly benefit both sets of disciplines.

Competition Recommendations:

10. What recommendations do you have to improve the competition?

To be honest, we thought this was a fantastic idea. Our only recommendation would be to repeat such an exercise with different types of datasets in the future (perhaps with more variables, different contexts, longer-term forecasting horizons, etc.). We think such an exercise would contribute to a much better understanding of what models, methods and approaches work best under what empirical conditions. All in all, we had a lot of fun participating in this competition, and look forward to future editions!

Pio Pio



Team leader: Juan Manuel Ramos-Goni

Team members: Mark Oppe and Oliver Rivero-Arias

Model Description:

1. Describe your choice of software and the reasons underlying your choice (e.g., Stata).

We conducted the modeling analysis in Stata version 14, as it was the software we all had enough experience with to get familiar with the data and run the statistical models and test and manage heteroscedasticity. The three members of the team had extensive experience programming in the package. We recognize that some models, such as generalized multinomial logit, had limitations in Stata (e.g., restriction of dimensions to 20), and other packages such as Nlogit could have been more suitable alternatives. However, too little information was provided about respondent characteristics to try other models of the generalized multinomial family.

2. Describe your choice of estimation technique and the reasons underlying your choice (e.g., Bayesian).

We have used a maximum likelihood estimation approach to modeling, as it is the standard estimation technique available to estimate the models we tested in Stata. We have used heteroscedastic models not only because they improve estimations but because the homoscedasticity assumption usually is not met for this type of data.

3. Describe your choice of functional form and the reasons underlying your choice (e.g., Logit).

We ran several functional forms during this exercise and started with the standard clogit as a reference case. We also tested cprobit, xtlogit, xtprobit, melogit, and meprobit. However, our initial intuition of preferring a heteroscedastic model led us to use hetprobit as the selected command. The clogithet was tested as well. However, this command does not provide the predicted probabilities (only xb), so calculating predictions became quite inefficient. So for practical reasons, we preferred hetprobit.

Note that probit and logit provide pretty similar predictions (logit being a bit better), so we assume that clogithet will produce very similar results to our prediction based on hetprobit.

4. Describe your choice of variables and the reasons underlying your choice (e.g., 20 effects-coded variables).

Our starting point was the interaction between the main effect (20 dummy-coded variables) and time. We then explored the inclusion of interaction effects, based on thee hypotheses: (1) Time units influence the choices; (2) the inclusion of immediate death influences the choices; and (3) the number of levels 4 and 5 on states A and B influences the choices. The final interactions included in our model are: (1) 4 dummies indicating time units (dead_year, month_dum, week_dum, day_dum); (2) a dummy representing immediate death (always coded as state A in the data; dead_dum); and (3) a dummy representing the number of levels 4 and 5, NR45_dif.

Modeling Recommendations:

5. Did you have difficulty modeling the 2 pair types (TTO pairs [quantity vs. quality] and efficient pairs [all attributes])? Did you have difficulty with the 4 temporal units (days, weeks, months, years)?

No. We just converted all times to the same units. However, our model suggests there are differences in choices explained by the time units, suggesting that the time proportionality assumption is not met for this type of data. Our model included corrections to be applied depending on how long the patient will be on the state.

6. Do you believe that you would have been able to predict choice probabilities better had you received data on the respondent characteristics as part of the exploratory dataset (e.g., age)? Why?

We believe that any statistical model should be decided on before one explores the data. We have enough evidence to suggest that modeling preference data is a complex process that involves not only people's attitudes to health and risks, but also the way they respond to and interact with the tool eliciting the preference. Therefore, any modelling technique should have recognized up front the different levels of heterogeneity involved (preference- and scale-heterogeneity). Models that recognize these issues require additional information on participants or choice-specific characteristics, and hence it would have been useful to have such data available during the modelling of the data.

7. Did you change your model's functional form or variables based on the estimation results (i.e., data mining)? If so, why and how? If not, why not?

Given that we had no prior analysis plan, our approach involved starting from a main-effect specification and proceeded to the inclusion of additional terms based on data exploration. We then tested several modelling approaches, each of them relaxing a particular assumption from the standard multinomial logit. We recognized that there was some element of data mining while working with the data. However, modelers should have previously thought about what could make sense to explore in the data. In other cases, we do not believe that success would be possible.

8. If your model wins, why do you believe it predicted better than the other models? If your model loses, why do you believe it did not predict better than the other models?

This is a very difficult question to answer, as we do not have any information about the workflow used by other team members. Our feeling is that we have achieved a very good MSE, but we are positive that other teams will arrive at similar figures using completely different strategies. This is what initiatives such as the Mount Hood challenge in diabetes have demonstrated (http://www.mthooddiabeteschallenge.com/).

9. Based on your expertise and experience, what are the primary econometric advances needed to improve predictive modeling (not design)?

We need to start recognizing the use of analysis plans when modeling health-preference data. Any such analysis should reflect previous experiences analyzing similar data (and we are already in that position), and models should incorporate aspects such as the fact that within a population there might be difference classes of preference heterogeneity, and aspects of learning and ordering effects when modelling the data.

Competition Recommendations:

10. What recommendations do you have to improve the competition?

We do not think a complex DCE using EQ-5D-5L states and duration was a good start for a first competition. We still have a lot to understand from a standard DCE to value health without duration.