EuroQol Working Paper Series

Number 16002
December 2016

ORIGINAL RESEARCH

Combining continuous and dichotomous responses in a hybrid model

Juan M. Ramos-Goñi[1]
Benjamin Craig[2]
Mark Oppe[3]
Ben van Hout[4]

[1] EuroQol Research Foundation, Rotterdam, the Netherlands
[2] University of South Florida, Tampa, USA
[3] EuroQol Research Foundation, Rotterdam, the Netherlands
[4] University of Sheffield, Sheffield, UK

EuroQol Research Foundation, Rotterdam, the Netherlands

www.euroqol.org

Abstract

Introduction: In survey research, continuous responses may represent a value, a lower or upper bound of a value, or a range of values (e.g., the value of my car is $10,000, is greater than or equal to $9,000 or is between $9,000 and $10,999). Dichotomous responses may represent an inequality in value (e.g., the value of my car is higher than the value of that car). Given a dataset with continuous and dichotomous responses, the hyreg command estimates the parameters of a hybrid regression model by maximizing a single likelihood function, namely the product of the likelihoods of continuous and dichotomous responses. Analogous to combining, for example, intreg and logit commands, this paper demonstrates the hyreg command using simulated data and includes an example of an econometric specification from health preference research.

Conflicts of interest: The authors have indicated they have no conflicts of interest with regard to the content of this article.

Keywords

TTO, DCE, Hybrid regression

Acknowledgements

We are grateful to the EuroQol Research Foundation for covering the fees of the authors in preparing this manuscript.

Juan Ramos-Goñi
EuroQol Research Foundation
Marten Meesweg 107
3068 AV Rotterdam
The Netherlands
E: jramos@euroqol.org

1 Introduction

In survey research, respondents are commonly asked to consider the location of objects along a scale using multiple tasks. For example, they may be asked to choose between 2 objects to express an inequality in value (e.g., car A is preferred to car B [A>B]). Such dichotomous responses facilitate the consideration of multiple differences in attributes and may mimic real world behaviour (i.e., action or inaction). Alternatively, respondents may be asked to place objects at points along a scale (e.g., I am willing to pay $10,000 for car A) or within intervals on a scale (e.g., I am willing to pay between $ 9,000 and $ 12,000 for car A). Respondents may be asked whether an object is located above or below a threshold (e.g., car A>$9,000; open interval). Unlike dichotomous responses, continuous responses (i.e., points or intervals along a scale) are more precise, but can be more cognitively burdensome for respondents as well as they require greater numeracy or understanding of labels. Whether a survey instrument captures the location of objects relative to other objects or at a point, within an interval, or above/below a threshold on a scale, survey researchers require an analytical approach that takes into account all available evidence. Notice that it is assumed high correlation between the different types of responses as all survey questions are related to the same objects, however, this assumption should be tested first. This paper introduces the hyreg command, which allows the estimation of a regression model with both continuous and dichotomous responses by maximizing a single likelihood function, namely the product of the likelihoods of dichotomous and continuous responses.

Like many innovations, this hybrid approach was borne out of necessity: specifically, Oppe and van Hout created an econometric approach, for modelling EQ-5D-5L valuation data that integrates dichotomous responses from discrete choice experiments (DCE; i.e., health A prefer to health B) with continuous responses from an iterative choice-based task, known as the time-trade off (TTO) [1,2]. Using an iterative process, the TTO task identifies the respondent's value of a health description in terms of years in full health (equivalent statement; i.e., health A for 10 years then die = full health for 8 years then die). Oppe and van Hout proposed to combine TTO and DCE responses in a single model, calling it the hybrid approach [2]. They suggest a maximum likelihood estimation of the product of the likelihood functions of a normal distribution for point observations (TTO responses) and a logistic model for dichotomous observations (DCE responses) based on the difference between the alternatives [3-5]. However, further review of the TTO responses revealed that their distribution was largely uniform with clustering on specific numbers of the iteration process, which complicated their interpretation [6].

During a scientific meeting in August 2014, Ramos-Goñi and Craig considered ignoring the equivalence statements of the TTO task and focusing on the iterative procedure that led up to the statement. The TTO choice-based process iteratively creates open and closed intervals (e.g., full health for 5 years < health state A for 10 years < full health for 8 years) as a means of narrowing in on the equivalence statement. As an exploratory analysis, these intervals were included as the dependent variable in the intreg command, which produced results with greater face validity than regular linear regression on the equivalence statements alone. Built from previous work on the hybrid approach [2-5], Ramos-Goñi and Craig decided to integrate the interval responses from the TTO with the dichotomous responses from the DCE under a common likelihood specification, which led to the development of the hyreg command.

The hyreg command further extends the hybrid approach to include 2 distributions (logistic and normal) and a multiplicative function of scaling (i.e., as hetprobit or intreg using het(#) option). The hyreg command also allows the dichotomous and continuous responses to have different distributions (logistic and normal) and have different independent variables to model scaling terms. Although originally developed for health preference research, the hyreg command can be used by anyone interested in combining continuous and dichotomous responses in a single maximum likelihood function to estimate the parameters of a regression model on a scale (e.g., sweetness, pain, wealth, value).

2 Description

Hyreg fits a hybrid model with both continuous and dichotomous responses by maximizing a single likelihood function.

3 Syntax

hyreg *depvar1* [*depvar2*] [*indepvars*] [if] [in] , datatype(*varname*) [interval contdist(*normal* | *logistic*) dichdist(*normal* | *logistic*) ll(#) ul(#) hetcont(*varlist*) hetdich(*varlist*) noconstant vce(oim | opg | robust | cluster *varname*) *maximize options*]

hyreg works in a similar way to most other Stata regression commands. Each observation includes one response described using one or two dependent variables (depvar1, depvar2) and one binary variable specified by datatype() (1 indicating that the response is continuous and 0 indicating that the response is dichotomous). A continuous response can be either a point or an interval (i.e., as for intreg). A dichotomous response is binary (0

or 1; i.e., as for probit). If the observations include only points and dichotomous responses, only one dependent variable is required (depvar1). If the observations also include interval responses, the hyreg command requires both the "interval" option and two dependent variables (depvar1, depvar2) indicating the boundaries of the interval. With the "interval" option, a point response is indicated when the two dependent variables have the same value (i.e., depvar1=depvar2). For open intervals (i.e., where either the left or right bounds are censored), the open end of the interval is represented by a missing value. In summary, the specification of depvar1 and depvar2 depend on the type of observation:

Type of observation		**depvar1**	**depvar2**
point observation	a = [a,a]	a	a
interval observation	[a,b]	a	b
left-censored observation	(-inf,b]	.	b
right-censored observation	[a,inf)	a	.
dichotomous observation	c	c	.

a, b ϵ]-∞,+ ∞[and c ϵ {0,1}

The contdist() and dichdist() options indicate the distributions for the continuous and dichotomous responses to be used in the maximum likelihood estimator. Point responses can have a lower limit (ll) and upper limit (ul; i.e., as tobit) and the scaling of each distribution may be associated with independent variables (e.g., heteroskedasticity). Therefore, the hyreg command includes distributional modifiers, namely ll(), ul(), hetcont(*varlist*), and hetdich(*varlist*). The default distributions are normal distribution for continuous responses and logistic distribution for dichotomous responses and do not include any modifiers.

4 Options

Model

datatype(*varname*) specifies the variable name containing the indicators of response type. An observation is 0 when a dichotomous response is present and 1 when a continuous response is present. datatype() is required.

interval is specified in the presence of a second dependent variable (depvar2). This second dependent variable allows the inclusion of intervals among the continuous responses (i.e., depvar1 is the lower bound, depvar2

is the upper bound) The open end of an interval is indicated by a missing value. With this option, a point response is indicated when the two dependent variables have the same value (i.e., depvar1=depvar2).

contdist(*normal* | *logistic*) specifies the distribution that the model fits over the continuous responses.

normal fits a normal distribution for continuous responses.

logistic fits a logistic distribution for continuous responses

dichdist(normal | logistic) specifies the distribution that the model fit over the dichotomous responses.

normal fits a normal distribution for dichotomous responses, as a probit model does.

logistic fits a logistic distribution for dichotomous responses, as a logistic model does.

ul(#) right-censoring limit such that all point responses greater than or equal to this limit are treated as censored.

ll(#) left-censoring limit such than all point responses less than or equal to this limit are treated as censored.

hetcont(*varlist***)** specifies the independent variables in the scaling function for the continuous distribution (i.e., lnsigma).

hetdich(*varlist***)** specifies the independent variables in the scaling function for the dichotomous distribution (i.e., lntheta).

noconstant suppresses the constant term (intercept) in the model of the scaled variable.

SE/Robust

vce(*vcetype*) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (**oim, opg**), that are robust to some kinds of misspecification (robust), that allow for intragroup correlation (***cluster*** *clustvar*).

Maximization

maximize options: difficult, technique(algorithm spec), iterate(#) , nolog, trace, gradient, showstep, hessian, showtolerance, tolerance(#), ltolerance(#), nrtolerance(#), nonrtolerance, and init(*init specs*).

5 Example

To illustrate how hyreg works we created a dataset of 1,000 respondents with 17 responses for each respondent. The 17 responses for each respondent include: 1 point response (Value of car A), 4 open interval responses (value of car A is higher or lower than a threshold), 5 closed interval responses (Value of car A is between X and Y) and 7 dichotomous responses representing inequalities (car A preferred to car B). This leads to 1,000 point responses. 2,000 left-censored intervals. 2,000 right-censored intervals. 5,000 closed intervals and 7,000 dichotomous responses. The values for continuous responses are scaled between 5-15 meaning thousands of

dollars and the values for dichotomous responses are 0-1 (1 if car B is chosen). The responses (N=17,000) has been stored in the file hyreg_data.dta.

. use hyreg_data.dta

. describe

```
  obs:        17,000
 vars:            37                          4 Mar 2016 20:14
 size:     2,312,000
```

variable name	storage type	display format	value label	variable label
id	long	%8.0g		
task	int	%8.0g		Task ID on study design
order_id	byte	%8.0g		Order of task presentation
colour	byte	%8.0g		Colour of the car (5 different colours)
wheels	byte	%8.0g		Type of the wheels (5 typess)
sport_4x4	byte	%8.0g		Type of car from sport to 4x4 type (5 types)
audio	byte	%8.0g		Type of audio system (5 types)
doors	byte	%8.0g		Type and number of doors (5 options)
B_colour	float	%9.0g		Colour of the car (5 different colours)
B_wheels	float	%9.0g		Type of the wheels (5 typess)
B_sport_4x4	float	%9.0g		Type of car from sport to 4x4 type (5 types)
B_audio	float	%9.0g		Type of audio system (5 types)
B_doors	float	%9.0g		Type and number of doors (5 options)
value	float	%8.0g		Responses (0-1 for inequalities and missing for continious values)
method	float	%9.0g		Binary variable indicating the type of method used to obtain the response
lower	double	%10.0g		Lower limit for the intervlas and inequality responses
upper	double	%10.0g		Upper limit for the intervlas and missing for inequality responses
colour2	float	%9.0g		Dummy (independent variable for the model)
colour3	float	%9.0g		Dummy (independent variable for the model)
colour4	float	%9.0g		Dummy (independent variable for the model)
colour5	float	%9.0g		Dummy (independent variable for the model)
wheels2	float	%9.0g		Dummy (independent variable for the model)
wheels3	float	%9.0g		Dummy (independent variable for the model)
wheels4	float	%9.0g		Dummy (independent variable for the model)
wheels5	float	%9.0g		Dummy (independent variable for the model)
sport_4x42	float	%9.0g		Dummy (independent variable for the model)
sport_4x43	float	%9.0g		Dummy (independent variable for the model)
sport_4x44	float	%9.0g		Dummy (independent variable for the model)
sport_4x45	float	%9.0g		Dummy (independent variable for the model)
audio2	float	%9.0g		Dummy (independent variable for the model)
audio3	float	%9.0g		Dummy (independent variable for the model)
audio4	float	%9.0g		Dummy (independent variable for the model)
audio5	float	%9.0g		Dummy (independent variable for the model)
doors2	float	%9.0g		Dummy (independent variable for the model)
doors3	float	%9.0g		Dummy (independent variable for the model)
doors4	float	%9.0g		Dummy (independent variable for the model)
doors5	float	%9.0g		Dummy (independent variable for the model)

Sorted by: id

The variable “method” indicates the type of response (0 for dichotomous responses; 1 for continuous responses).

. tab method

```
. tab method
```

Binary variable indicating the type of method used to obtain the response	Freq.	Percent	Cum.
0	7,000	41.18	41.18
1	10,000	58.82	100.00
Total	17,000	100.00	

The data for each respondent is as follows: for the continuous responses, the independent variables represent the description of car A (colour to doors). For dichotomous responses, we also include variables describing the alternative (car B; B_colour to B_doors).

```
. list id-upper if id ==2590
```

	id	task	order_id	colour	wheels	sport~x4	audio	doors	B_colour	B_wheels	B_spor~4	B_audio	B_doors	value	method	lower	upper
1.	2590	53	1	2	2	4	3	4	.	.	.	.	.	.	1	10	11
2.	2590	50	2	2	4	5	5	3	.	.	.	.	.	.	1	12	12.5
3.	2590	83	3	1	1	2	1	1	.	.	.	.	.	.	1	5	5.25
4.	2590	52	4	1	1	4	2	5	.	.	.	.	.	.	1	6.5	7
5.	2590	54	5	4	2	1	1	5	.	.	.	.	.	.	1	7	7.5
6.	2590	49	6	1	3	1	2	2	.	.	.	.	.	.	1	.	7.5
7.	2590	56	7	4	5	4	1	3	.	.	.	.	.	.	1	.	10
8.	2590	55	8	3	5	3	3	2	.	.	.	.	.	.	1	10	.
9.	2590	51	9	5	1	1	5	2	.	.	.	.	.	.	1	10	.
10.	2590	86	10	5	5	5	5	5	.	.	.	.	.	.	1	13.5	13.5
11.	2590	113	11	5	2	1	1	1	1	1	4	3	1	1	0	1	.
12.	2590	135	12	5	4	5	5	5	3	5	5	3	5	1	0	1	.
13.	2590	122	13	2	1	3	5	4	4	1	3	2	1	1	0	1	.
14.	2590	27	14	3	4	1	3	2	2	4	4	4	5	0	0	0	.
15.	2590	105	15	2	4	5	2	3	4	5	1	2	5	0	0	0	.
16.	2590	169	16	1	1	4	4	5	3	2	1	1	5	0	0	0	.
17.	2590	11	17	3	5	2	1	1	4	2	5	5	1	0	0	0	.

Prior to analysis, all variables representing the attributes of cars A are recoded as dummy variables for continuous responses. In case of dichotomous responses, the dummy variables of car B are subtracted from the variables of car A so that the recoded dummy variables represent the differences between car A and B.

The default specification includes a normal distribution for continuous responses and the logistic distribution for dichotomous responses. For purposes of simulation, the constant term was dropped. To incorporate the open and closed intervals, the hybrid command must include the “interval” option and a second dependent variable (i.e., depvar2). In this case, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) interval nocons nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.344763	.0884481	15.20	0.000	1.171408	1.518119
colour3	1.024789	.0984525	10.41	0.000	.8318259	1.217753
colour4	2.312743	.0966544	23.93	0.000	2.123303	2.502182
colour5	2.488422	.096671	25.74	0.000	2.298951	2.677894
wheels2	1.611344	.084098	19.16	0.000	1.446515	1.776173
wheels3	1.267405	.0979788	12.94	0.000	1.07537	1.45944
wheels4	1.952684	.0962394	20.29	0.000	1.764058	2.141309
wheels5	2.224468	.0888176	25.05	0.000	2.050388	2.398547
sport_4x42	1.692613	.08951	18.91	0.000	1.517176	1.868049
sport_4x43	1.25203	.0942327	13.29	0.000	1.067337	1.436723
sport_4x44	2.083887	.0956769	21.78	0.000	1.896363	2.27141
sport_4x45	2.069083	.0882375	23.45	0.000	1.896141	2.242025
audio2	1.596118	.0838017	19.05	0.000	1.431869	1.760366
audio3	1.338447	.0986938	13.56	0.000	1.145011	1.531883
audio4	2.303966	.0946902	24.33	0.000	2.118377	2.489555
audio5	2.929835	.0952562	30.76	0.000	2.743136	3.116533
doors2	2.094111	.0885908	23.64	0.000	1.920476	2.267746
doors3	1.96096	.0985397	19.90	0.000	1.767826	2.154094
doors4	3.071502	.0900714	34.10	0.000	2.894965	3.248039
doors5	3.217783	.088312	36.44	0.000	3.044694	3.390871
lnsigma						
_cons	1.253981	.0095759	130.95	0.000	1.235213	1.27275
lntheta						
_cons	-.8315329	.0344434	-24.14	0.000	-.8990408	-.764025

```
      1000  continuous uncensored observations
      2000  continuous left-censored observations
      2000  continuous right-censored observations
      5000  continuous interval observations
      7000  dichotomous observations
```

With a normal distribution for the dichotomous responses, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) contdist(normal) dichdist(normal) interval nocons nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.361203	.0896599	15.18	0.000	1.185472	1.536933
colour3	1.022867	.0993353	10.30	0.000	.8281736	1.217561
colour4	2.339319	.0979142	23.89	0.000	2.147411	2.531227
colour5	2.504011	.0976231	25.65	0.000	2.312673	2.695349
wheels2	1.623739	.0849729	19.11	0.000	1.457195	1.790283
wheels3	1.249896	.0986605	12.67	0.000	1.056525	1.443267
wheels4	1.933073	.0974855	19.83	0.000	1.742005	2.124141
wheels5	2.208389	.0898828	24.57	0.000	2.032222	2.384556
sport_4x42	1.700799	.0901446	18.87	0.000	1.524119	1.877479
sport_4x43	1.246483	.0949004	13.13	0.000	1.060482	1.432484
sport_4x44	2.083483	.0964581	21.60	0.000	1.894428	2.272537
sport_4x45	2.059563	.0890736	23.12	0.000	1.884982	2.234144
audio2	1.616608	.084741	19.08	0.000	1.450518	1.782697
audio3	1.32147	.0995103	13.28	0.000	1.126434	1.516507
audio4	2.314139	.0954957	24.23	0.000	2.126971	2.501307
audio5	2.933666	.0959325	30.58	0.000	2.745641	3.12169
doors2	2.103319	.0894827	23.51	0.000	1.927936	2.278702
doors3	1.968084	.0994225	19.80	0.000	1.773219	2.162948
doors4	3.069516	.0906838	33.85	0.000	2.891779	3.247253
doors5	3.203245	.0896107	35.75	0.000	3.027611	3.378879
lnsigma						
_cons	1.253046	.0095754	130.86	0.000	1.234279	1.271814
lntheta						
_cons	1.363679	.0322494	42.29	0.000	1.300471	1.426886

```
        1000  continuous uncensored observations
        2000  continuous left-censored observations
        2000  continuous right-censored observations
        5000  continuous interval observations
        7000  dichotomous observations
```

With a logistic distribution for the continuous responses, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) contdist(logistic) dichdist(logistic) interval nocons nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.263469	.088078	14.34	0.000	1.090839	1.436099
colour3	.901853	.0974724	9.25	0.000	.7108105	1.092896
colour4	2.206804	.0954919	23.11	0.000	2.019643	2.393965
colour5	2.416781	.0961424	25.14	0.000	2.228345	2.605216
wheels2	1.612092	.0836299	19.28	0.000	1.448181	1.776004
wheels3	1.202755	.096541	12.46	0.000	1.013538	1.391972
wheels4	1.885458	.0961765	19.60	0.000	1.696956	2.073961
wheels5	2.228747	.0881642	25.28	0.000	2.055949	2.401546
sport_4x42	1.694819	.0899525	18.84	0.000	1.518515	1.871123
sport_4x43	1.2378	.0927731	13.34	0.000	1.055968	1.419631
sport_4x44	2.049346	.0947613	21.63	0.000	1.863618	2.235075
sport_4x45	2.083945	.0872743	23.88	0.000	1.912891	2.254999
audio2	1.59081	.0831156	19.14	0.000	1.427906	1.753713
audio3	1.240593	.0976139	12.71	0.000	1.049273	1.431913
audio4	2.22149	.0940198	23.63	0.000	2.037215	2.405766
audio5	2.921267	.0947456	30.83	0.000	2.735569	3.106965
doors2	2.166209	.087646	24.72	0.000	1.994426	2.337992
doors3	1.950307	.0972055	20.06	0.000	1.759787	2.140826
doors4	3.025481	.0884988	34.19	0.000	2.852026	3.198935
doors5	3.259518	.087192	37.38	0.000	3.088624	3.430411
lnsigma						
_cons	.6971123	.0109073	63.91	0.000	.6757343	.7184902
lntheta						
_cons	-.8290934	.0349098	-23.75	0.000	-.8975153	-.7606715

```
        1000  continuous uncensored observations
        2000  continuous left-censored observations
        2000  continuous right-censored observations
        5000  continuous interval observations
        7000  dichotomous observations
```

With a logistic distribution for the continuous responses and a normal distribution for the dichotomous responses, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) contdist(logistic) dichdist(normal) interval nocons nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.280671	.0894655	14.31	0.000	1.105322	1.45602
colour3	.8992594	.0982985	9.15	0.000	.7065979	1.091921
colour4	2.23168	.0967706	23.06	0.000	2.042013	2.421347
colour5	2.431784	.0971365	25.03	0.000	2.2414	2.622168
wheels2	1.62347	.0845193	19.21	0.000	1.457815	1.789124
wheels3	1.18437	.0972124	12.18	0.000	.9938369	1.374903
wheels4	1.864428	.0972932	19.16	0.000	1.673736	2.055119
wheels5	2.212532	.0891569	24.82	0.000	2.037788	2.387276
sport_4x42	1.702845	.0903023	18.86	0.000	1.525855	1.879834
sport_4x43	1.232745	.0933047	13.21	0.000	1.049871	1.415619
sport_4x44	2.048329	.0954312	21.46	0.000	1.861287	2.235371
sport_4x45	2.075412	.0880446	23.57	0.000	1.902848	2.247976
audio2	1.611453	.0838582	19.22	0.000	1.447094	1.775813
audio3	1.224691	.0982763	12.46	0.000	1.032073	1.417309
audio4	2.232013	.0948568	23.53	0.000	2.046097	2.417929
audio5	2.926141	.0953487	30.69	0.000	2.739261	3.113022
doors2	2.175888	.0884085	24.61	0.000	2.002611	2.349165
doors3	1.956895	.0979885	19.97	0.000	1.764841	2.148949
doors4	3.021881	.0890481	33.94	0.000	2.847349	3.196412
doors5	3.245708	.0884435	36.70	0.000	3.072362	3.419054
lnsigma						
_cons	.6961352	.0109066	63.83	0.000	.6747586	.7175119
lntheta						
_cons	1.361181	.0327107	41.61	0.000	1.297069	1.425293

```
        1000  continuous uncensored observations
        2000  continuous left-censored observations
        2000  continuous right-censored observations
        5000  continuous interval observations
        7000  dichotomous observations
```

With a normal distribution for the continuous responses and a logistic distribution for the dichotomous responses, but using heteroscedasticity in both types of responses, the hybrid model estimates are as follows:

```
. hyreg lower upper colour2-doors5 , datatype(method) interval nocons nolog hetcont(colour2-doors5) hetdich(colour2-doors5)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
model						
colour2	1.215438	.097968	12.41	0.000	1.023424	1.407452
colour3	.9361072	.104181	8.99	0.000	.7319161	1.140298
colour4	2.180632	.1056835	20.63	0.000	1.973496	2.387767
colour5	2.362725	.103163	22.90	0.000	2.160529	2.564921
wheels2	1.445028	.0919531	15.71	0.000	1.264803	1.625252
wheels3	1.303646	.1042452	12.51	0.000	1.099329	1.507963
wheels4	1.793598	.1098829	16.32	0.000	1.578231	2.008965
wheels5	2.087382	.1000534	20.86	0.000	1.891281	2.283483
sport_4x42	1.564673	.1012316	15.46	0.000	1.366263	1.763083
sport_4x43	1.206338	.106827	11.29	0.000	.9969612	1.415715
sport_4x44	1.985284	.1095693	18.12	0.000	1.770532	2.200036
sport_4x45	2.032859	.0948608	21.43	0.000	1.846935	2.218782
audio2	1.320055	.0992176	13.30	0.000	1.125592	1.514518
audio3	1.260709	.1089929	11.57	0.000	1.047086	1.474331
audio4	2.092974	.1011976	20.68	0.000	1.89463	2.291318
audio5	2.684734	.1051208	25.54	0.000	2.478701	2.890767
doors2	2.435524	.123007	19.80	0.000	2.194434	2.676613
doors3	2.448527	.1164185	21.03	0.000	2.220351	2.676703
doors4	3.355396	.1015937	33.03	0.000	3.156276	3.554516
doors5	3.590764	.1046951	34.30	0.000	3.385565	3.795962
lnsigma						
colour2	-.0832428	.0352389	-2.36	0.018	-.1523097	-.0141758
colour3	-.0246814	.035188	-0.70	0.483	-.0936487	.0442858
colour4	-.0234911	.0401845	-0.58	0.559	-.1022513	.0552692
colour5	.0251924	.0349982	0.72	0.472	-.0434028	.0937876
wheels2	-.1091662	.0340023	-3.21	0.001	-.1758095	-.0425229
wheels3	-.0830598	.0404313	-2.05	0.040	-.1623037	-.0038158
wheels4	.0169463	.0365636	0.46	0.643	-.054717	.0886095
wheels5	-.0086818	.0349379	-0.25	0.804	-.0771587	.0597952
sport_4x42	-.1038585	.0357219	-2.91	0.004	-.1738721	-.0338448
sport_4x43	-.1569476	.0378076	-4.15	0.000	-.2310492	-.0828461
sport_4x44	-.0657407	.0371338	-1.77	0.077	-.1385217	.0070403
sport_4x45	-.0460983	.0335391	-1.37	0.169	-.1118338	.0196372
audio2	-.1309788	.0335284	-3.91	0.000	-.1966931	-.0652644
audio3	-.0523165	.0380126	-1.38	0.169	-.1268198	.0221868
audio4	.0679624	.0328714	2.07	0.039	.0035356	.1323891
audio5	-.0473537	.0349757	-1.35	0.176	-.1159047	.0211974
doors2	-.3984228	.0414845	-9.60	0.000	-.4797309	-.3171147
doors3	-.3664519	.0416811	-8.79	0.000	-.4481453	-.2847585
doors4	-.3148747	.0382326	-8.24	0.000	-.3898091	-.2399402
doors5	-.2867316	.0378698	-7.57	0.000	-.360955	-.2125082
_cons	1.642797	.0339145	48.44	0.000	1.576325	1.709268
lntheta						
colour2	.023478	.0899211	0.26	0.794	-.1527642	.1997202
colour3	-.0629817	.0820764	-0.77	0.443	-.2238484	.097885
colour4	-.0534122	.0902596	-0.59	0.554	-.2303178	.1234933
colour5	.1208891	.0844548	1.43	0.152	-.0446393	.2864175
wheels2	-.0317514	.0774907	-0.41	0.682	-.1836304	.1201276
wheels3	.0524146	.0735013	0.71	0.476	-.0916453	.1964746
wheels4	.0642982	.0870393	0.74	0.460	-.1062956	.234892
wheels5	.1947116	.0747065	2.61	0.009	.0482895	.3411337
sport_4x42	-.0981628	.0623218	-1.58	0.115	-.2203112	.0239856
sport_4x43	.0271457	.0672912	0.40	0.687	-.1047426	.159034
sport_4x44	.0568761	.070923	0.80	0.423	-.0821306	.1958827
sport_4x45	.1017443	.073639	1.38	0.167	-.0425855	.2460741
audio2	-.2327513	.0786215	-2.96	0.003	-.3868465	-.078656
audio3	-.0611286	.0756355	-0.81	0.419	-.2093715	.0871143
audio4	.2391177	.069731	3.43	0.001	.1024474	.3757881
audio5	.1768521	.0680425	2.60	0.009	.0434911	.310213
doors2	.1420926	.0798455	1.78	0.075	-.0144017	.2985869
doors3	-.3092774	.0716293	-4.32	0.000	-.4496682	-.1688867
doors4	-.2833589	.0629347	-4.50	0.000	-.4067087	-.1600091
doors5	-.0915457	.0704827	-1.30	0.194	-.2296893	.0465979
_cons	-.8620898	.0375803	-22.94	0.000	-.9357458	-.7884338

```
        1000  continuous uncensored observations
        2000  continuous left-censored observations
        2000  continuous right-censored observations
        5000  continuous interval observations
        7000  dichotomous observations
```

6 Saved results

hyreg stores the following in e():

Scalars

e(rank)	rank of e(V)
e(N)	number of observations
e(ic)	number of iterations
e(k)	number of parameters
e(k_eq)	number of equations in e(b)
e(k_dv)	number of dependent variables
e(converged)	1 if converged, 0 otherwise
e(rc)	return code
e(N_clust)	number of clusters
e(ll)	log likelihood
e(k_eq_model)	number of equations in overall model test
e(df_m)	model degrees of freedom
e(chi2)	chi-squared
e(p)	p-value for model chi-squared test

Macros

e(cmd)	used command
e(chi2type)	Wald type of model chi-squared test
e(opt)	type of optimization
e(predict)	program used to implement predict
e(vcetype)	title used to label Std. Err.
e(clustvar)	name of cluster variable
e(vce)	vcetype specified in vce()
e(user)	name of likelihood-evaluator program
e(ml_method)	type of ml method
e(technique)	maximization technique
e(which)	max or min; whether optimizer is to perform maximization or minimization
e(depvar)	names of dependent variable
e(properties)	b V

Matrices

e(b)	coefficient vector
e(V)	variance-covariance matrix of the estimators
e(ilog)	iteration log (up to 20 iterations)
e(gradient)	gradient vector
e(V_modelbased)	model-based variance

Functions

e(sample)	marks estimation sample

7 Methods and formulas

hyreg fits, by maximum likelihood, a hybrid regression model, $x\beta + \varepsilon$, where β denotes the vector of model coefficients, x denotes the independent variables of the model and ε represents the error term. The dependent

variable of the model, y_j, depends on the type of observation: y_j is a continuous point response for observations j ∈ C and y_j is a dichotomous response for observations j ∈D. Caution is warranted when merging responses from different techniques into a single estimator [5]. The variance of the continuous responses may not be equal to the variance found in dichotomous responses. The continuous and dichotomous error terms may even have entirely different distributions. Oppe and van Hout used a normal distribution for continuous responses and a logistic distribution for dichotomous responses [2], obtaining the following log-likelihood formula:

(Formula 1)

$$lnL = -\frac{1}{2} * \sum_{j \in C} \left\{ \ln(2\pi\sigma^2) + \left(\frac{y_j - x\beta}{\sigma}\right)^2 \right\} + \sum_{j \in D} \left\{ ln \left(\frac{1}{1 + e^{(-x\beta')}}\right) * y_j + ln \left(\frac{e^{(-x\beta')}}{1 + e^{(-x\beta')}}\right) * (1 - y_j) \right\}$$

This formula includes only point and dichotomous responses and serves as the default specification for the hyreg command. Noting that the logit coefficients of the dichotomous model, β' may not be on the same scale as the coefficients of the continuous model, β, due to distributional differences, they introduced a proportional rescaling parameter θ, such that β' = β /θ:

(Formula 2)

$$lnL = -\frac{1}{2} * \sum_{j \in C} \left\{ \ln(2\pi\sigma^2) + \left(\frac{y_j - x\beta}{\sigma}\right)^2 \right\} + \sum_{j \in D} \left\{ ln \left(\frac{1}{1 + e^{(-x\beta/\theta)}}\right) * y_j + ln \left(\frac{e^{(-x\beta/\theta)}}{1 + e^{(-x\beta/\theta)}}\right) * (1 - y_j) \right\}$$

For the hyreg command, this log-likelihood was extended to allow for left-censored (L), right-censored (R), and closed intervals (I) obtaining the formula (i.e., as intreg):

(Formula 3)

$$lnL = -\frac{1}{2} * \sum_{j \in C} \left\{ \ln(2\pi\sigma^2) + \left(\frac{y_j - x\beta}{\sigma}\right)^2 \right\} + \sum_{j \in L} ln \left(\Phi \left(\frac{y_{Lj} - x\beta}{\sigma} \right) \right) + \sum_{j \in R} ln \left(\Phi \left(\frac{-(y_{Rj} - x\beta)}{\sigma} \right) \right) + \sum_{j \in I} ln \left(\Phi \left(\frac{y_{2j} - x\beta}{\sigma} \right) - \Phi \left(\frac{y_{1j} - x\beta}{\sigma} \right) \right) + \sum_{j \in D} \left\{ ln \left(\frac{1}{1 + e^{(-x\beta/\theta)}}\right) * y_j + ln \left(\frac{e^{(-x\beta/\theta)}}{1 + e^{(-x\beta/\theta)}}\right) * (1 - y_j) \right\}$$

Alternatively, the distribution of the error terms may be the same for the continuous and dichotomous responses (i.e., normal-probit or logistic-logit), obtaining the following 2 log-likelihood formulae respectively:

(Formula 4)

$$lnL = -\frac{1}{2} * \sum_{j \in C} \left\{ \ln(2\pi\sigma^2) + \left(\frac{y_j - x\beta}{\sigma}\right)^2 \right\}$$

$$+ \sum_{j \in L} ln \left(\Phi \left(\frac{y_{Lj} - x\beta}{\sigma} \right) \right)$$

$$+ \sum_{j \in R} ln \left(\Phi \left(\frac{-(y_{Rj} - x\beta)}{\sigma} \right) \right)$$

$$+ \sum_{j \in I} ln \left(\Phi \left(\frac{y_{2j} - x\beta}{\sigma} \right) - \Phi \left(\frac{y_{1j} - x\beta}{\sigma} \right) \right)$$

$$+ \sum_{j \in D} \left\{ \ln \left(\Phi \left(\frac{-x\beta}{\theta} \right) \right) * (1 - y_j) + ln \left(\Phi \left(\frac{x\beta}{\theta} \right) \right) * y_j \right\}$$

(Formula 5)

$$lnL = \sum_{j \in C} ln \left(\frac{e^{\frac{-(y_j - x\beta)}{\sigma}}}{\sigma * \left(1 + e^{\frac{-(y_j - x\beta)}{\sigma}} \right)^2} \right)$$

$$+ \sum_{j \in L} ln \left(\frac{1}{1 + e^{\frac{-(y_{Lj} - x\beta)}{\sigma}}} \right)$$

$$+ \sum_{j \in R} ln \left(1 - \frac{1}{1 + e^{\frac{-(y_{Rj} - x\beta)}{\sigma}}} \right)$$

$$+ \sum_{j \in I} ln \left(\left(\frac{1}{1 + e^{\frac{-(y_{2j} - x\beta)}{\sigma}}} \right) - \left(\frac{1}{1 + e^{\frac{-(y_{1j} - x\beta)}{\sigma}}} \right) \right)$$

$$+ \sum_{j \in D} \left\{ ln \left(\frac{1}{1 + e^{(-x\beta/\theta)}} \right) * y_j + ln \left(\frac{e^{(-x\beta/\theta)}}{1 + e^{(-x\beta/\theta)}} \right) * (1 - y_j) \right\}$$

Technical note:

For implementation purpose, the hyreg command estimates ln(σ) and ln(θ), instead of σ and θ directly. These parameters, ln(σ) and ln(θ), may be modelled using separate regressions to allow for heteroskedasticity (i.e., as hetprobit or using the het option of the intreg command).

7.1 The specific case of TTO and DCE data

The EQ-5D-5L valuation datasets include point responses from a TTO task and paired comparison responses from a DCE task. These tasks have 3 potential implications, which serve to demonstrate the capabilities of the *hyreg* command. First, the rescaling parameter for the TTO responses may be proportionally associated with the EQ-5D-5L attributes (i.e., heteroskedasticity). In other words, worse health implies greater potential variability in value. Second, the point responses equal to -1 were recorded as -1, not allowing for responses less than -1 (left-censoring; i.e., as *tobit*). Third, independent variables in the paired comparison responses represent additive differences between the attributes of the alternatives, x_A-x_B (i.e., as *logit*) [6].

The three implications are demonstrated as modifications to formula 3:

(Formula 6)

$$lnL = -\frac{1}{2} * \sum_{j \in C'} \left\{ \ln(2\pi(e^{z\gamma})^2) + \left(\frac{y_j - x\beta}{e^{z\gamma}}\right)^2 \right\} + \sum_{j \in L'} ln\left(\Phi\left(\frac{-1 - x\beta}{e^{z\gamma}}\right)\right) + \sum_{j \in D} \left\{ ln\left(\frac{1}{1 + e^{(-(x_A - x_B)\beta/\theta)}}\right) * y_j + ln\left(\frac{e^{(-(x_A - x_B)\beta/\theta)}}{1 + e^{(-(x_A - x_B)\beta/\theta)}}\right) * (1 - y_j) \right\}$$

Where $\sigma = e^{z\gamma}$ (i.e., $\ln(\sigma)=z\gamma$), C' represents TTO responses greater than -1, L' represents TTO responses of -1, and x_A and x_B represent the attributes of alternatives A and B in the paired comparisons.

Alternatively, some analysts may be accustomed to maximizing conditional log-likelihood functions to fit models of dichotomous responses (i.e., as *clogit*). Unlike Formula 6, these functions include separate observations for each alternative (no differences) and assemble the observations in groups [7]. However, this approach is not directly amenable to the integration with continuous responses, particularly normal distributions. For the modelling of a scaled variable using continuous and dichotomous responses, *hyreg* provides a common framework for normal and logistic distributional specifications, separates the distributional specifications by response type (e.g., normal-logit), allows censoring of points and lower and upper bounds, and can relax homoskedasticity assumptions.

8 Acknowledgements

We are grateful to the EuroQol Research Foundation for covering the fees of the authors in preparing this manuscript.

9 References

1 Oppe M, Devlin NJ, van Hout B, Krabbe PF, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17(4):445-53.

2 Oppe M, van Hout B. The optimal hybrid: experimental design and modeling of a combination of TTO and DCE. EuroQol Group Proceedings. 2013. Available at: http://www.euroqol.org/uploads/media/EQ2010_-_CH03_-_Oppe_-_The_optimal_hybrid_-_Experimental_design_and_modeling_of_a_combination_of_TTO_and_DCE.pdf. Accessed October 11, 2014.

3 Rowen D, Brazier J, Van Hout B. A comparison of methods for converting DCE values onto the full health-dead QALY scale. Med Decis Making. 2015 Apr;35(3):328-40. doi: 10.1177/0272989X14559542. Epub 2014 Nov 14.

4 Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. Med Care. 2014 Dec 17. [Epub ahead of print]

5 Craig BM, Busschbach JJ. The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. Popul Health Metr. 2009 Jan 13;7:3. doi: 10.1186/1478-7954-7-3. Available at: http://ww.pophealthmetrics.com/content/7/1/3

6 Craig BM, Runge SK, Rand-Hendriksen K, Ramos-Goñi JM, Oppe M. Learning and satisficing: an analysis of sequence effects in health valuation. Value Health. 2015 Mar;18(2):217-23. doi: 10.1016/j.jval.2014.11.005. Epub 2015 Feb 2.

7 Train, K. Discrete Choice Methods with Simulation. Second edition. Cambridge University Press, 2009.

About the authors

Juan M. Ramos-Goñi is Senior researcher at the EuroQol Research Foundation, The Netherlands.

Benjamin M. Craig is Associate Professor of Economics at the University of South Florida, USA.

Mark Oppe is Senior researcher at the EuroQol Research Foundation, The Netherlands.

Ben van Hout is Professor at the University of Sheffield, UK.