基于离散选择模型的结直肠癌筛查偏好研究

苏锦华 *
*中国人民大学统计学院

{2017201620}@*ruc.edu.cn*

***Index Terms*—癌症筛查偏好；DCE；多元logit模型**

I. 相关研究

为更好地开展癌症筛查工作，政策制定者需了解目标人群的筛查偏好，找到并权衡影响其参加癌症筛查的属性（如筛查效果），这是设计与实施科学合理筛查方案的关键。

陈述性偏好（stated preference）方法利用离散选择实验（discrete choice experiments，DCEs）研究目标人群癌症筛查偏好，系统分析影响其参加癌症筛查的重要因素[1]。

DCE对被调查对象的分析基于随机效用理论（random utility theory）[2]。在构建回归模型时，将筛查方案是否被受访者选中作为因变量，癌症筛查属性作为自变量。由于因变量是哑变量的属性，logit或probit模型常用于估计目标人群对各个癌症筛查属性的效用值，从而得到筛查方案总的效用值。

有关DCE的研究大多采用随机效果probit模型（random effects probit）、条件logit模型（conditional logit）、多项式Logit模型（multinomial logit）等经典模型；近年来，嵌套logit模型（nested logit）、混合logit模型（mixed logit，考虑了受访者的选择异质性[**milte2014cognitive**]（preference heterogeneity））、广义多项式logit模型（generalised multinomial logit，同时考虑了受访者的选择异质性和规模异质性[**kjaer2008preference**]（scale heterogeneity））也逐渐成为研究人员开展DCE研究所采用的模型。

II. 数据

A. 数据处理

回收网络问卷525份，问卷设置了10道个人特质问题，3大类方案，其中选择方案是3种属性的组合。由于对不同个体，不同医院来说，三类选择方案中风险降低程度、复查频次是不存在唯一的客观值，通过改变风险、频次得到18种假设情景供受访者选择。

将每一种假设情景作为一条数据，每份问卷实际上可分成18条供模型训练的数据，总计9449条训练数据，数据变量取值与含义如下。

B. 数据描述

对9449条数据的各属性分别统计频次直方图。总体上三套方案选择人数相近，乙状结肠镜的筛查方案稍微低于另外两套方案。个人特征属性的数据分布情况各异：年龄分布上50-60岁人数居多，对健康的重视程度普遍较高，直结肠检查比例大致80%，慢性病比例70%，受教育程度平均为高中教育,选择复查间隔较短的人数较多，有住院经历的占60%，80%的人群年收入在5万以下,65%的人群居住在城市，55%的受访者尚未退休,65%的受访者为女性,大部分受访者更青睐风险降低程度较大的方案。

TABLE I
数据变量解释

sex	性别:0表示男，1代表女
age	年龄:0 3表示50周岁到70周岁（5岁一挡）,4表示70周岁以上
income	年收入:0表示五万一下，1表示5-10万，2表示10万以上
region	居住地区:0表示乡镇，1表示市区
edu	学历:0表示初中及以下，1表示高中及中专，2表示本科及大专，3表示研究生以上
work	就业状态:1表示工作，0表示不工作
retire	退休状态:1表示退休，0表示尚未退休
chronic	是否有慢性病史:1表示是，0表示否
check	是否做过结直肠检查:1表示是，0表示否
hospital	是否有过住院经历:1表示是，0表示否
attention	健康关注程度:0 4健康关注程度逐步提高
risk_dec	筛查方案风险降低百分比（%）
frequency	筛查方案每隔x年需复查
pain	筛查方案检查机器深入肠道的长短（cm）
choice	筛查方案选择:0表示粪便潜血试验，1表示乙状结肠镜，2表示全结肠镜

III. 模型

本文选择多元逻辑回归来探究不同因素对筛查方案选择偏好的影响。因素分为个人特征因素与选择情景因素，本文构建了普通多元logit模型和带交互项的多元logit模型。普通多元logit模型假设各因素间相互独立互不影响，而带交互项的多元logit模型考虑个人特征对偏好的复合影响。

为了使模型具有较好的解释性，本文通过前进法依次选取对模型影响最显著的变量纳入模型，去除共线性较高使得模型训练难以收敛的变量。

A. 多元*logit*模型

1) 模型公式*:* U是基于随机效用假设的效用，n代表第n个样本，j代表第j个选择方案，s代表选择情景，在本文中代表风险和频次不同的18个假设情景。V为多元logit模型估计的效用值，选择第j个方案概率是所有方案效用的softmax值。

$$U_{njs} = \beta_j x_{njs} + \epsilon_{njs} \tag{1}$$

$$V_{njs} = \beta_j x_{njs} \tag{2}$$

$$Pr_{njs}(\beta) = \frac{e^{\beta_j x_{njs}}}{\sum_{i=1}^{J} e^{\beta_i x_{nis}}} \tag{3}$$

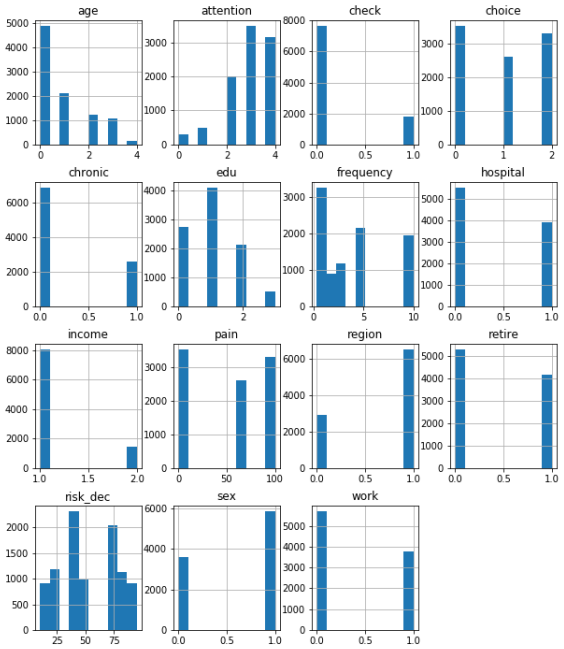

Fig. 1. 数据直方图

2) 前进法选择变量*:* 本文选择了前进法选择前10个变量拟合多元logit模型。在假设情景数据中，频次的影响最显著，风险降低的影响不显著，pain由于在不同情景中未发生变化，存在与choice较高的共线性未被选入。收入、性别、教育、工作情况对偏好的影响较显著，而年龄、就医经历、病史、对健康的重视程度则对偏好的影响不显著。

TABLE II
模型一前进法选择结果

$frequency >> intercept >> income >> sex >> edu >> work >> retire >> check >> hospital >> age >> chronic >> region >> risk\_dec >> attention$

B. 带交互项的多元*logit*模型

1) 模型公式*:* 在普通多元logit模型中，beta是需要拟合客观数值。现实情况中不同的个人特征对方案偏好有影响，本文模型二假设beta是个人特征属性的线性加性函数，将公式（5）带入公式（4）可以得到个人属性与情景属性的交互项。

$$U_{njs} = \beta_j^{'}(character)situation_{njs} + \epsilon_{njs}^{'} \tag{4}$$

$$\beta_j^{'}(character) = \beta_{j0}^{'} + \beta_{j1}^{'}sex + \beta_{j2}^{'}age + ... \tag{5}$$

```
==============================================================================
Dep. Variable:                 choice   No. Observations:                 9449
Model:                        MNLogit   Df Residuals:                     9429
Method:                           MLE   Df Model:                           18
Date:                Tue, 24 Mar 2020   Pseudo R-squ.:                  0.2449
Time:                        08:47:50   Log-Likelihood:                -7782.6
converged:                       True   LL-Null:                       -10307.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
  choice=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
frequency      0.7359      0.018     40.926      0.000       0.701       0.771
intercept     -1.6397      0.160    -10.226      0.000      -1.954      -1.325
income        -0.5676      0.111     -5.133      0.000      -0.784      -0.351
sex           -0.3133      0.066     -4.740      0.000      -0.443      -0.184
edu            0.0927      0.043      2.170      0.030       0.009       0.176
work          -0.0975      0.101     -0.962      0.336      -0.296       0.101
retire         0.1576      0.097      1.628      0.104      -0.032       0.347
check         -0.3623      0.090     -4.031      0.000      -0.538      -0.186
hospital       0.0792      0.067      1.181      0.238      -0.052       0.211
age            0.0633      0.032      1.971      0.049       0.000       0.126
------------------------------------------------------------------------------
  choice=2       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
frequency      0.7855      0.018     43.533      0.000       0.750       0.821
intercept     -2.8482      0.149    -19.106      0.000      -3.140      -2.556
income         0.4956      0.095      5.212      0.000       0.309       0.682
sex           -0.1482      0.066     -2.252      0.024      -0.277      -0.019
edu            0.2307      0.042      5.544      0.000       0.149       0.312
work          -0.4721      0.096     -4.920      0.000      -0.660      -0.284
retire        -0.4964      0.093     -5.340      0.000      -0.679      -0.314
check         -0.0785      0.087     -0.898      0.369      -0.250       0.093
hospital       0.2545      0.066      3.846      0.000       0.125       0.384
age            0.0704      0.032      2.226      0.026       0.008       0.132
==============================================================================
```

Fig. 2. 多元logit模型

2) 前进法选择变量*:* 本文选择了前进法选择前15个变量拟合带交互项的多元logit模型。频次依旧是最显著的影响因素，其他被选入的重要变量均为交互项，剩余的较重要的非交互项只有受教育程度，说明不同个人特征的确对筛查偏好的选择有重要影响。含有收入、对健康的重视程度、是否工作与退休、受教育程度、年龄的交互项较显著，说明以上个人特质属性对筛查偏好有重要影响。

TABLE III
模型二前进法选择结果

$frequency >> intercept >> income * risk_dec >> income * frequency >> attention * risk\_dec >> attention * frequency >> retire * risk_dec >> retire * frequency >> work * risk\_dec >> work * frequency >> edu >> edu * frequency >> age * risk\_dec >> age * frequency >> sex * pain >> chronic * risk\_dec >> region >> chronic * frequency >> region * frequency >> hospital * risk\_dec >> chronic * pain >> region * risk\_dec >> check * frequency >> check * risk\_dec >> check * pain$

IV. 结论

模型有三个选项，以choice=0对基准，分别对剩余两个变量进行logit回归，其系数正负与大小均以choice=0的系数为0作为基准。

V. 模型一拟合结果解读

频率拟合系数均为正向高度显著，且choice=2¿choice=1系数，说明复查间隔越大，第三个方案的边际对数效用提升最大，其次是第二个。可以理解当复查间隔越长，二三方案更可能被选择。

choice=1收入影响是负向显著，choice=2收入影响是正向显著，说明当收入越高，越可能选择方案三，其次是方案一，选择方案二的概率减少。

MNLogit Regression Results

Dep. Variable:	choice	No. Observations:	9449
Model:	MNLogit	Df Residuals:	9419
Method:	MLE	Df Model:	28
Date:	Tue, 24 Mar 2020	Pseudo R-squ.:	0.8138
Time:	09:30:44	Log-Likelihood:	-1919.6
converged:	False	LL-Null:	-10307.
Covariance Type:	nonrobust	LLR p-value:	0.000

choice=1	coef	std err	z	P>|z|	[0.025	0.975]
frequency	3.2829	0.204	16.063	0.000	2.882	3.684
intercept	-11.9407	0.506	-23.580	0.000	-12.933	-10.948
income*risk_dec	0.0650	0.007	9.845	0.000	0.052	0.078
income*frequency	-0.7803	0.098	-7.989	0.000	-0.972	-0.589
attention*risk_dec	0.0324	0.003	11.886	0.000	0.027	0.038
attention*frequency	-0.3573	0.041	-8.738	0.000	-0.437	-0.277
retire*risk_dec	0.0435	0.008	5.196	0.000	0.027	0.060
retire*frequency	-0.5776	0.133	-4.359	0.000	-0.837	-0.318
work*risk_dec	0.0678	0.008	8.280	0.000	0.052	0.084
work*frequency	-0.7447	0.127	-5.884	0.000	-0.993	-0.497
edu	-0.7230	0.133	-5.446	0.000	-0.983	-0.463
edu*frequency	0.1940	0.044	4.451	0.000	0.109	0.279
age*risk_dec	0.0135	0.002	5.441	0.000	0.009	0.018
age*frequency	-0.1142	0.037	-3.060	0.002	-0.187	-0.041
sex*pain	0.1195	0.007	16.675	0.000	0.105	0.134

choice=2	coef	std err	z	P>|z|	[0.025	0.975]
frequency	5.2919	0.224	23.631	0.000	4.853	5.731
intercept	-30.5320	0.892	-34.233	0.000	-32.280	-28.784
income*risk_dec	0.2418	0.011	21.890	0.000	0.220	0.263
income*frequency	-2.1045	0.122	-17.309	0.000	-2.343	-1.866
attention*risk_dec	0.0487	0.003	15.886	0.000	0.043	0.055
attention*frequency	-0.4930	0.044	-11.272	0.000	-0.579	-0.407
retire*risk_dec	0.0691	0.009	7.518	0.000	0.051	0.087
retire*frequency	-0.8952	0.140	-6.407	0.000	-1.169	-0.621
work*risk_dec	0.1013	0.009	11.152	0.000	0.083	0.119
work*frequency	-1.0616	0.135	-7.876	0.000	-1.326	-0.797
edu	-1.5767	0.182	-8.684	0.000	-1.933	-1.221
edu*frequency	0.3037	0.047	6.413	0.000	0.211	0.397
age*risk_dec	0.0185	0.003	6.427	0.000	0.013	0.024
age*frequency	-0.1581	0.041	-3.903	0.000	-0.237	-0.079
sex*pain	0.1353	0.007	18.578	0.000	0.121	0.150

Fig. 3. 带交互项的多元logit模型

性别的影响系数均为负向显著，且choice=1负值程度更大，说明女性相比男性更倾向与选择方案一，对方案二的拒绝程度比方案三更大。

教育的影响系数均为正向显著，且choice=2系数数值更大，说明受教育程度高的人更倾向于选择方案三，其次是方案二。

是否工作与是否退休的拟合系数的均仅有choice=2显著，都为负无论工作还是退休都更愿意选择方案一而非方案三。

是否经历过直结肠检查的系数仅有choice=1显著为负，说明经历过直结肠检查的人对方案三既不偏好也不厌恶，而在方案一与方案二中更愿意选择方案一。

是否有住院史的系数仅有choice=2显著为正，说明有住院史的人在方案一与方案三中更愿意选择方案三，对方案一和方案二则没有明显差别。

年龄系数均为正向显著，choice=2数值更大，说明年龄越大，越倾向于选择方案三，其次是方案二。

VI. 模型二拟合结果解读

频率拟合系数规律仍然同模型一解读，差别是方案三和方案二的边际偏好差值更大了。

收入与风险降低的交互系数均为正向显著，且choice=2数值更大，说明收入不同的人群对风险降低的效用感是不同，收入更高的人群对风险降低的偏好更大。

收入与频率拟合的系数可能存在与频率系数的共线性，缺乏解读意义，正是因为这种共线性，模型二的系数数值明显大于模型一。

对健康重视程度更高的人群、退休人群、年龄较高的人群均表现出和高收入人群相同的偏好变化，即更倾向于风险降低程度更高的方案，更偏好方案三而非方案二。而对频率的重视程度也呈现负值，也是由于与频率系数的共线性。

值得注意的是模型一中未被选入pain属性在模型二中与性别属性的交互呈现显著，说明女性比男性对深入直结肠的长度更加在意。