计量经济学-作业10-虚拟应变量

😀 [redacted]

P247 习题2

2 R. Amatya [⊖]利用尼泊尔 1 145 名年龄介于 35～44 岁之间的已婚女性的数据，估计出了如下有关生育控制的 logit 模型：

$$\widehat{L:Pr(D_i=1)} = -4.47 + 2.03WN_i + 1.45ME_i$$
$$(0.36) \qquad (0.14)$$

式中，D_i 代表第 i 位女性是否采用过生育控制措施，为虚拟变量，如果采用过则为 1，否则为 0；WN_i 代表第 i 位女性是否不想要小孩，为虚拟变量，如果不想则为 1，否则为 0；ME_i 代表第 i 位女性了解的生育控制方法的数量。

a. 请解释参数 WN 和 ME 的理论意义。如果是线性概率模型的话，答案会有什么不同呢？

b. 斜率参数估计值的符号、大小、显著程度是否符合预期？为什么？

c. 在这个方程中，常数项的显著程度理论上应该是什么样的？

d. 若要改变模型设定，应该怎么改呢？为什么？

回答

a. 解释理论意义：

- `WN`:在ME保持不变的情况下，不想要小孩的女性 会比 想要小孩的女性 采用生育控制措施的**可能性（概率）**增加50.75% *(2.03*0.5*0.5)*
- `MN`:在MN保持不变的情况下，女性了解的生育控制方法每增加一个单位，女性采用生育控制措施的**可能性（概率）**增加36.25% *(1.45*0.5*0.5)*
- **如果是线性概率模型**，那么单个斜率参数表示：当其他变量不变时，每增加1单位解释变量对女性采用已知生育控制方法的概率的影响。

b. **符号：**符合预期。两者皆为正数，而根据经验理论，如果不想要小孩，那么更可能采取控制生育的方法减少其怀孕；而如果越多的了解生育控制方法，那么其选择更多，也更有可能采取生育控制措施。

大小：基本符合预期，大小处在合理的程度。

显著性：建立假设 $H_0:\beta \leq 0; H_A:\beta > 0$，$t_1 = \frac{2.03}{0.36} = 5.638$，$t_2 = \frac{1.45}{0.14} = 10.35$，$t_1 > t_c, t_2 > t_c$。且备择假设和预期符号一致，因此可以拒绝原假设。因此两个参数都具有显著性。符合预期，但是WN的斜率参数的t值比ME的斜率参数t值大（即 $\hat{\beta}_2$ 比 $\hat{\beta}_1$ 更加显著）稍微在预期之外。

c. 常数项的显著性程度理论上**没有意义**，因为常数项对被解释变量的影响微不足道

d. 模型可能存在遗漏变量，可以增加变量“**女性受教育程度**”。受教育程度和他们是否采取生育控制往往比较相关，因为在教育中他们会更加明确知道采取生育的危害以及在什么情况下必须采取生育控制。

P247 习题4

4 回到表 13-1 所示的女性进入劳动力市场的数据，考虑将第 i 位女性的年龄 A_i 加入方程。在确定预期符号和函数形式的时候需要特别小心，因为年龄对（女性）是否参与劳动力市场的预期影响很难把握。例如，一部分女性在婚后退出了劳动力市场，而另外一部分女性即便在抚养孩子的时候仍继续工作。还有一部分女性结婚后就待在家里，不去工作，当小孩到了上学年龄时，她们又回到劳动力市场。例如，马尔科姆·科恩（Malcolm Cohen）等发现女性年龄对女性是否参与劳动力市场的影响很小，除非年龄达到 65 岁或更高并很可能面临退休。[⊖]最终结果为，在女性参与劳动力市场的模型中，年龄似乎是一个在理论上不相干的变量。由此，作为一个可能的预期，把年龄作为虚拟变量，如果第 i 位女性年龄大于或等于 65 岁为 1，否则为 0。

a. 观察表 13-1 的数据集。加入一个解释变量（当第 i 位女性年龄大于或等于 65 岁时虚拟变量取值为 1，否则为 0）会出什么问题？

b. 实践估计自己的线性概率和 logit 方程，检验年龄（A_i）是女性劳动力参与模型相关变量的可能性。用表 13-1 的数据估计。然后在我们的标准要求下用你的方程与文中的版本对比。解释为什么认为年龄是（或不是）相关变量（提示：要计算 $\overline{R}_p^2$）。

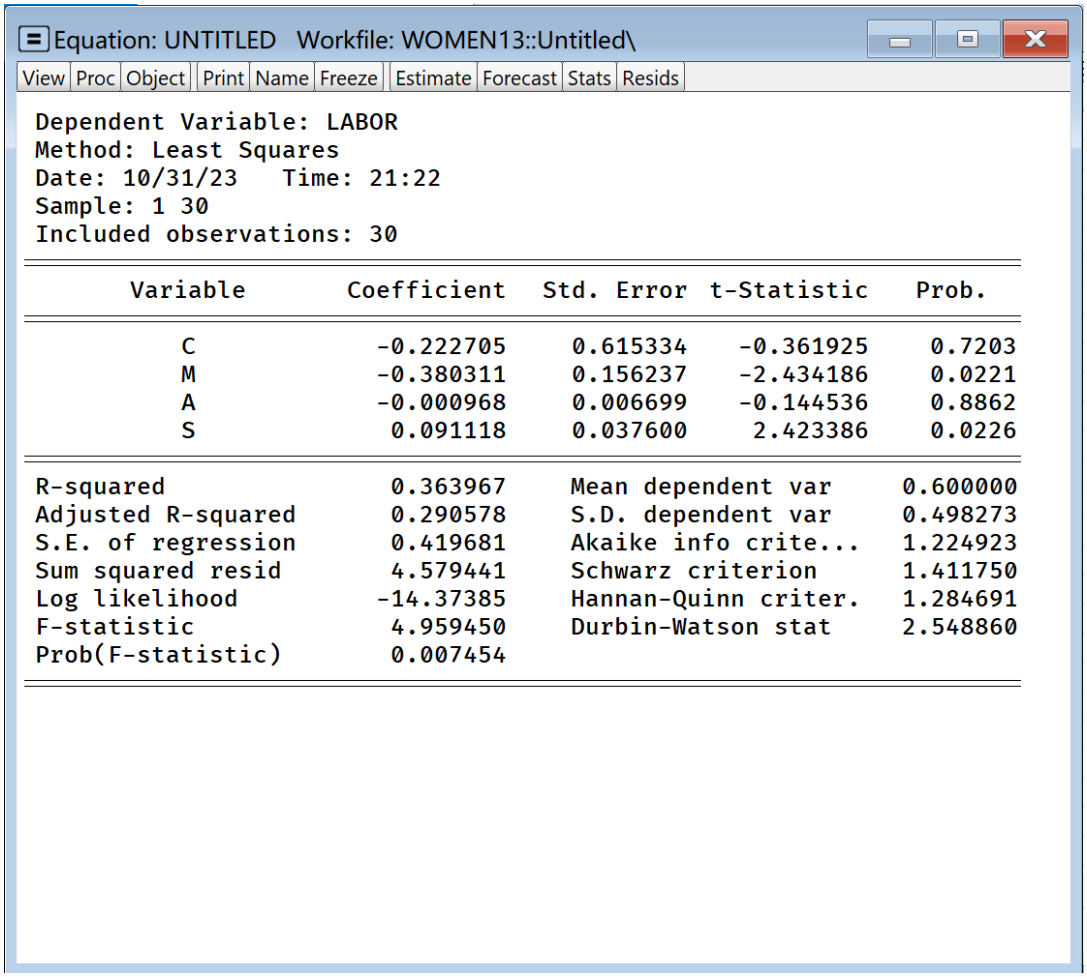
回答

a. 样本中仅有2位女性年龄超过65岁，且这两位女性都不是劳动力。如果加入一个是否退休（年龄65）的解释变量，那么会发现当年龄超过65岁时全部样本均为非劳动力，而不存在是劳动力的样本。**这就产生了一个近奇矩阵，导致我们无法估计这个方程。**

b.

i. **线性概率**

输入“ls labor c m a s”,使用Eviews软件进行回归分析



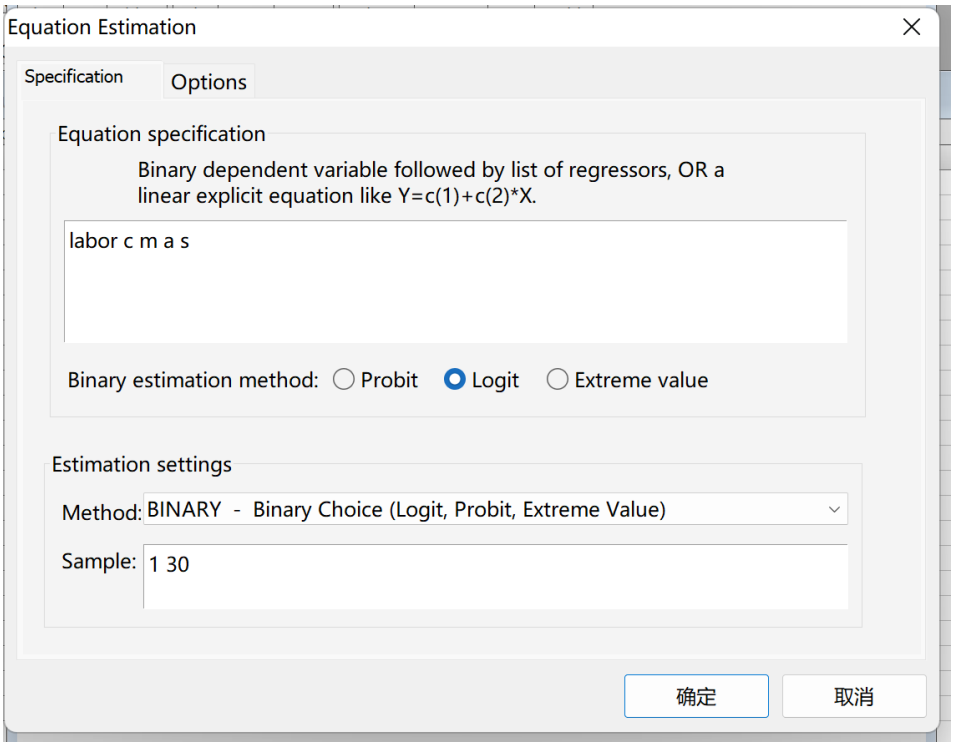
即回归方程计为：

$$\widehat{D_i} = -0.22 - 0.38M_i - 0.001A_i - 0.09S_i$$
$$(0.16) \quad (0.007) \quad (0.04)$$
$$t = -2.43 \quad -0.14 \quad 2.42$$
$$N = 30 \quad \overline{R^2} = 0.29 \quad \overline{R_P^2} = 0.806$$

其中 $\overline{R_P^2}$ 是被解释变量的值（1和0）被正确解释的百分比的平均值

ii. Logit模型

进入Eviews选择 `Binary estimation method:Logit`



得出回归结果为——

Equation: UNTITLED Workfile: WOMEN13::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LABOR
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 10/31/23 Time: 22:08
Sample: 1 30
Included observations: 30
Convergence achieved after 6 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-5.266331	4.120739	-1.278006	0.2012
M	-2.608477	1.199746	-2.174192	0.0297
A	-0.009888	0.038964	-0.253777	0.7997
S	0.674164	0.321575	2.096444	0.0360

McFadden R-squared	0.344013	Mean dependent var	0.600000
S.D. dependent var	0.498273	S.E. of regression	0.406527
Akaike info criterion	1.149641	Sum squared resid	4.296878
Schwarz criterion	1.336467	Log likelihood	-13.24461
Hannan-Quinn criter.	1.209408	Deviance	26.48922
Restr. deviance	40.38070	Restr. log likeli...	-20.19035
LR statistic	13.89148	Avg. log likelihood	-0.441487
Prob(LR statistic)	0.003057		

Obs with Dep=0	12	Total obs	30
Obs with Dep=1	18		

得出结果

$$\widehat{L:Pr(D_i=1)} = -5.27 - 2.61M_i - 0.01A_i - 0.67S_i$$
$$(1.20) \quad (0.04) \quad (0.32)$$
$$t = -2.17 \quad -0.25 \quad 2.10$$
$$N = 30 \qquad \overline{R_P^2} = 0.76$$

P247 习题5

5 2008 年，Goldman 和 Romley [⊖] 通过分析大洛杉矶地区医疗保险覆盖的 117 家医院的8 721位肺炎患者的数据来研究医疗需求，发现患者在选择医院时，临床质量（用较低的肺炎死亡率衡量）相对于其他因素来说，作用较小。

观察 Goldman 和 Romley 的样本的一个子集发现：499 名患者要么选择 UCLA 医疗中心，要么选择附近的 Cedars Sinai 医疗中心。通常情况下，经济学家会认为价格在这样的选择中起主要作用，但对于有医疗保险的患者来说，无论选择哪家医院，他们的支出都几乎一致。相反，诸如患者的住宅与医院的距离、患者的年龄以及患者的收入等因素成为重要的潜在影响因素。

$$\widehat{L:Pr(D_i=1)} = 4.41 - 0.38DISTANCE_i - 0.072INCOME_i - 0.29OLD_i \quad (13\text{-}23)$$
$$(0.05) \qquad (0.036) \qquad (0.31)$$
$$N = 499 \qquad \overline{R}_p^2 = 0.66 \qquad \text{迭代次数：8 次}$$

式中，D_i 代表第 i 位患者是不是选择 Cedars Sinai 医疗中心，为虚拟变量，如果选择它，则为 1，而选择 UCLA 医疗中心则为 0；$DISTANCE_i$ 代表第 i 位患者的居住地与 Cedars Sinai 医疗中心的距离（根据邮政编码）减去该患者的居住地与 UCLA 医疗中心的距离（单位：英里）；$INCOME_i$ 代表第 i 位患者的收入（用邮政编码所在区域内的平均收入衡量，单位：千美元）；OLD_i 代表第 i 位患者是否超过了 75 岁，为虚拟变量，如果超过了，为 1，否则为 0。

a. 对 $DISTANCE$ 的参数做出假设，并在 5%的显著水平下进行检验。

b. 请仔细说明变量 $DISTANCE$ 的参数估计值的经济意义，即距离每改变 1 单位对选择 Cedars Sinai 医疗中心的概率造成的影响。

c. 请仔细思考 $DISTANCE$ 的定义，为什么要将 $DISTANCE$ 定义为两段距离之差，而不将与 Cedars Sinai 医疗中心的距离和与 UCLA 医疗中心的距离作为两个不同的解释变量呢？

d. 这个数据集可以在网上找到（www. pearsonhighered. com /Studenmund），文件名为 HOSPITAL13。将数据导入电脑，采用 Stata 或你的电脑中的其他回归程序估计方程的线性概率形式。在你估计出的方程中，$DISTANCE$ 的参数估计值是什么样的？哪个模型更好？请给出解释。

e. 现在，构建一个斜率虚拟变量 $OLD * DISTANCE$，将其加入方程（13-17）并估计出新的 logit 模型。为什么要引入这个特别的斜率虚拟变量？对这个斜率虚拟变量的参数做出假设，并在 5%的显著水平下进行检验。方程（13-17）与新的具有斜率虚拟变量的 logit 方程哪个更好？请给出说明。（选做）

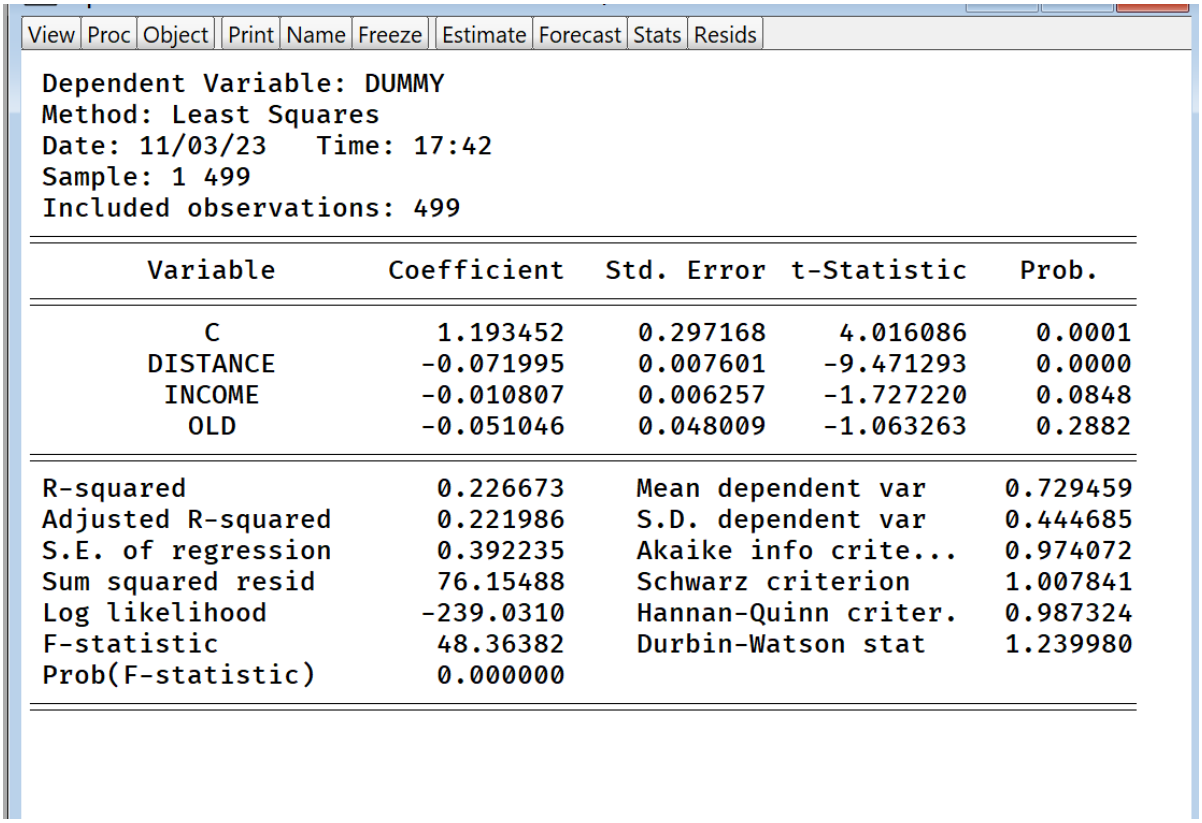
回答

a. DISTANCE预期符号为负 *（因为距离越远，越不选择这个医院）*。因此在5%显著性水平下建立假设：$H_0: \beta \geq 0; H_A: \beta < 0$。自由度为499-4-495，而 $t_c = 1.65$，$t_1 = \dfrac{-0.38}{0.05} = -7.6$，|-7.6|>1.65，即 $|t_1| > t_c$，而备择假设符号和预期一致，因此，**可以拒绝原假设。**

b. 经济意义为：在其他变量保持不变的前提下，患者的居住地与Cedars Sinai医疗中心的距离 与该患者居住地与UCLA医疗中心的距离的差值每多1英里，患者最终选择Cedars Sinai医疗中心的概率就会下降9.5%（*0.38*0.5*0.5*）。

c. 因为大多数患者关心的是到两家医院的相对距离，而不是到医院的个人距离的绝对值。

d. 采用线性模型拟合出的结果为：



可以看到，线性概率模型中的**DISTANCE系数为-0.072**。

采用logit模型的（13-23）更好，因为当因变量是虚拟应变量时，我们需要避免估计为线性概率模型，原因是这样拟合出来因变量的范围是无限的，不符合实际情况。

e. 新增OLD*DISTANCE这个交叉项的意义是：由于老年患者的行动能力有限，老年患者可能比年轻患者更有可能尽量缩短去医院的距离。因此，**预期交叉项符号为负数**，做出假设

$H_0 : \beta \geq 0; H_A : \beta < 0$

使用Eviews拟合的结果为：

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: DUMMY
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 11/03/23 Time: 17:43
Sample: 1 499
Included observations: 499
Convergence achieved after 4 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	4.253543	1.732766	2.454771	0.0141
DISTANCE	-0.106331	0.097925	-1.085833	0.2776
INCOME	-0.070361	0.036294	-1.938606	0.0525
OLD	-0.216105	0.287708	-0.751129	0.4526
OLD*DISTANCE	-0.327814	0.108219	-3.029163	0.0025

McFadden R-squared	0.209793	Mean dependent var	0.729459
S.D. dependent var	0.444685	S.E. of regression	0.392078
Akaike info criterion	0.942679	Sum squared resid	75.94027
Schwarz criterion	0.984890	Log likelihood	-230.1985
Hannan-Quinn criter.	0.959244	Deviance	460.3971
Restr. deviance	582.6287	Restr. log likeli...	-291.3143
LR statistic	122.2316	Avg. log likelihood	-0.461320
Prob(LR statistic)	0.000000		

Obs with Dep=0	135	Total obs	499
Obs with Dep=1	364		

可以看到OLD*DISTANCE的z分数大于临界值，并且北泽假设与预期符号一致，所以我们可以拒绝原假设。

这个新的方程更好一些，从理论上讲，这个交叉项具备经济理论意义，由于老年患者的行动能力有限，老年患者可能比年轻患者更有可能尽量缩短去医院的距离也符合实际。方程也具备显著性和一定的拟合优度。引入参数后，各个系数的偏误也不大。