计量经济学第三次作业

Karry

1.假定你通过调查收集到了关于工资、教育程度、经验以及性别等个人信息。除此之外，你还询问了有关个人吸食大麻的情况（询问被调查者“在过去的一个月之内吸食了多少次大麻”）。然后，建立了以下的模型：

$$log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 femal + u$$

因此，$100\beta_1$表示（每月）每增加一次大麻的吸食会导致工资的近似百分比变化。

a) 请修正模型使得你可以直接检验吸食大麻对男女生工资会有不同的影响；并说明在新建的模型中如何检验“吸食大麻对男女生工资具有相同的影响”的假设？

> 答：为了能够直接检验吸食大麻对男女工资的影响，可以将模型修正为：
>
> $$log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 femal + \beta_6 femal * usage + u$$
>
> 解释：我在原本的模型中引入了性别和吸食大麻情况的交叉项即：$femal * usage$ 这一项能够表示吸食大麻和性别之间的交互作用。
>
> 检验“吸食大麻对男女生工资具有相同的影响”这一假设，针对修正后的模型而言，表述为检验原假设 $H_0: \beta_6 = 0$ 可以发现要检验的参数只有一个，所以直接对 β_6 进行 t 检验即可。

b) 如果你认为不应该直接使用“吸食大麻的次数”作为解释变量，而最好将吸食大麻的变量表示成一个四值的分类变量：从不吸食，轻度吸食（每月1-5次）、中度吸食（每月6-10次）、重度吸食（每月10次以上）。那么此时的模型应该做出什么样的调整来使得你可以考察不同程度的大麻吸食对工资的影响？

> 答：应该将模型调整为：
>
> $$log(wage) = \beta_0 + \beta_1 usage_{l1} + \beta_2 usage_{l2} + \beta_3 usage_{l3} + \beta_4 usage_{l4} + \beta_5 educ + \beta_6 exper + \beta_7 exper^2 + \beta_8 femal + u$$
>
> 其中 $usage_{li}$ 的含义如表 1.

变量名	变量含义
$usage_{l1}$	从不吸食<属于此类则该变量为 1 否则为 0 以下同理>
$usage_{l2}$	轻度吸食（每月1-5次）
$usage_{l3}$	中度吸食（每月6-10次）
$usage_{l4}$	重度吸食（每月10次以上）

表 1. 有关 usage 的变量含义表

回归后通过 β_1、β_2、β_3、β_4 的大小即可考察出不同程度的大麻吸食对工资的影响。

c) 在b) 部分建立的模型中，如何检验大麻吸食对工资没有影响的原假设？

答：大麻吸食对工资没有影响的原假设，在 b) 部分的模型中可以表示为虚拟假设：

$$H_0 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$$

这时我们就可以将题干中的原模型看作 b) 部分模型的有监督模型，b) 部分模型为无监督模型，通过以下公式计算 F 统计量的大小来进行检验即可。

$$F = \frac{(SSR_r - SSR_{ur})/(df_r - df_{ur})}{SSR_{ur}/df_{ur}}$$

d) 如果吸食大麻与饮酒之间是正相关的，并且饮酒与工资之间是负向关联（在控制教育、经验以及性别之后）。那么，如果遗漏反映个人饮酒情况的变量会对β_1的估计产生什么样的影响？

答：会造成 β_1 的估计值偏小，解释如下：

我们可以将这个问题抽象为工资(wage)对吸食大麻(usage)和饮酒(alcohol)进行回归，假设遗漏了个人饮酒状况模型估计结果为 $\tilde{wage} = \tilde{\beta_0} + \tilde{\beta_1} usage$ 添加个人饮酒状况的模型估计结果为 $\hat{wage} = \hat{\beta_0} + \hat{\beta_1} wage + \hat{\beta_2} alcohol$

可以知道遗漏饮酒状况所估计出来的 $\tilde{\beta_1}$ 与不遗漏估计出的 $\hat{\beta_1}$ 满足如下关系：

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\hat{\delta}_1$$

由题意可知：$\hat{\beta}_2 < 0, \hat{\delta}_1 > 0$，因此 $\tilde{\beta}_1 < \hat{\beta}_1$

即遗漏反映个人饮酒情况的变量会对造成 β_1 的估计值偏小。

2.a) 估计以下的模型：

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u$$

其中hsize代表高中时的班级人数（单位是百人）。报告回归的基本结果。并检验hsize平方项的系数是否显著？

答：所作回归的基本结果如图1.所示：（其中 hsizesq 为 hisizesq 的平方项）

```
reg sat hsize hsizesq
```

Source	SS	df	MS		
				Number of obs	= 4,137
				F(2, 4134)	= 15.93
Model	614822.097	2	307411.048	Prob > F	= 0.0000
Residual	79759024.2	4,134	19293.4263	R-squared	= 0.0076
				Adj R-squared	= 0.0072
Total	80373846.3	4,136	19432.7481	Root MSE	= 138.9

sat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	19.81446	3.990666	4.97	0.000	11.99061	27.63831
hsizesq	-2.130606	.549004	-3.88	0.000	-3.206949	-1.054263
_cons	997.9805	6.203448	160.88	0.000	985.8184	1010.143

图 1. sat 对班级人数和班级人数的平方回归结果

可以从回归结果中发现：hsize 平方项（hsizesq）的系数在 1% 的水平上显著异于 0 也就是说在 1% 的显著水平上水平上 hsize平方项的系数在统计学上是显著的！

b) 基于以上的估计结果，计算最优的班级规模。

答：由以上结果可以得到估计的模型如下

$$\hat{sat} = 997.98 + 19.81 hsize - 2.13 hsize^2$$

由二次函数性质可知 $\hat{sat}$ 在 $\frac{19.81}{2\times 2.13} = 4.65$ 处取得最大值。

也即最优班级规模为 465 人

c) 如果使用log(sat)作为被解释变量，再次估计模型，并计算最优的班级规模，请问与b）中的最优规模存在很大的差异吗？

> 答：以 log(sat) 作为被解释变量，估计模型变为：
>
> $$log(sat) = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u$$
>
> 该模型的估计结果如图2. 所示

```
reg log_sat hsize hsizesq
```

Source	SS	df	MS	Number of obs	=	4,137
				F(2, 4134)	=	16.19
Model	.614405203	2	.307202602	Prob > F	=	0.0000
Residual	78.4287724	4,134	.018971643	R-squared	=	0.0078
				Adj R-squared	=	0.0073
Total	79.0431776	4,136	.01911102	Root MSE	=	.13774

log_sat	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	.0196029	.0039572	4.95	0.000	.0118445	.0273612
hsizesq	-.0020872	.0005444	-3.83	0.000	-.0031546	-.0010199
_cons	6.896029	.0061515	1121.03	0.000	6.883969	6.908089

图 2. log(sat) 对班级人数和班级人数的平方回归结果

> 同 b）由二次函数性质可知 $\hat{sat}$ 在 $\frac{0.0196}{2\times 0.0021} = 4.68$ 处取得最大值。
>
> 也即最优班级规模为 468 人 可以看出此时的最优规模和 b）中只相差 3 人（不到1%）差距很小！

d) 估计以下模型：

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u$$

请问 β_6 估计值的含义。基于估计的结果，该估计值统计上是否显著？

> 答：该模型的估计结果如图3. 所示

```
reg colgpa hsize hsizesq hsperc sat female athlete
```

Source	SS	df	MS	Number of obs	=	4,137
				F(6, 4130)	=	284.59
Model	524.819305	6	87.4698842	Prob > F	=	0.0000
Residual	1269.37637	4,130	.307355053	R-squared	=	0.2925
				Adj R-squared	=	0.2915
Total	1794.19567	4,136	.433799728	Root MSE	=	.5544

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0568543	.0163513	-3.48	0.001	-.0889117	-.0247968
hsizesq	.0046754	.0022494	2.08	0.038	.0002654	.0090854
hsperc	-.0132126	.0005728	-23.07	0.000	-.0143355	-.0120896
sat	.0016464	.0000668	24.64	0.000	.0015154	.0017774
female	.1548814	.0180047	8.60	0.000	.1195826	.1901802
athlete	.1693064	.0423492	4.00	0.000	.0862791	.2523336
_cons	1.241365	.0794923	15.62	0.000	1.085517	1.397212

图 3.

β_6 的估计值含义为：在其他条件（hsize、hsperc、sat、female）都给定的情况下，是运动员的学生，预计其大学GPA（colgpa）相对于不是运动员的学生约高出 0.17

可以从回归结果中看出，β_6 的估计值在 1% 的显著性水平上都是统计学显著的。

f） 如果在问题d) 的模型中允许“是不是运动员”对成绩产生的影响在男女生中有不同，建立对应的模型，并运用估计的结果检验“女性运动员与女性非运动员的成绩没有差异”的原假设。

答：引入运动员(athlete)和性别(female)的交叉项，进而建立对应的模型如下：

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + \beta_7 athlete * female + u$$

此模型的估计结果如图4. 所示

```
reg colgpa hsize hsizesq hsperc sat female athlete fem_p_ath
```

Source	SS	df	MS		
Model	524.821272	7	74.9744674	Number of obs	= 4,137
Residual	1269.3744	4,129	.307429015	F(7, 4129)	= 243.88
				Prob > F	= 0.0000
				R-squared	= 0.2925
				Adj R-squared	= 0.2913
Total	1794.19567	4,136	.433799728	Root MSE	= .55446

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0568006	.0163671	-3.47	0.001	-.0888889	-.0247124
hsizesq	.0046699	.0022507	2.07	0.038	.0002573	.0090825
hsperc	-.0132114	.000573	-23.06	0.000	-.0143349	-.012088
sat	.0016462	.0000669	24.62	0.000	.0015151	.0017773
female	.1546151	.0183122	8.44	0.000	.1187133	.1905168
athlete	.1674185	.0484877	3.45	0.001	.0723564	.2624806
fem_p_ath	.0076921	.0961748	0.08	0.936	-.1808623	.1962466
_cons	1.241575	.0795453	15.61	0.000	1.085623	1.397526

图 4.

其中 fem_p_ath 即 athlete * female 由于这是第一次接触到此类变量的生成方法，所以将所使用的代码展示如下：

```
gen fem_p_ath=athlete if female==1
replace fem_p_ath=0 if female==0
```

检验“女性运动员与女性非运动员的成绩没有差异”的原假设，对应于本模型中即检验原假设

$$H_0:\beta_6=0 \text{ 且 } \beta_7=0$$

令$\beta_6=0,\beta_7=0$构建约束模型：

$$colgpa=\beta_0+\beta_1 hsize+\beta_2 hsize^2+\beta_3 hsperc+\beta_4 sat+\beta_5 female+u$$

此模型的回归结果如图5. 所示

```
reg colgpa hsize hsizesq hsperc sat female
```

Source	SS	df	MS	Number of obs	=	4,137
				F(5, 4131)	=	337.09
Model	519.906875	5	103.981375	Prob > F	=	0.0000
Residual	1274.2888	4,131	.308469813	R-squared	=	0.2898
				Adj R-squared	=	0.2889
Total	1794.19567	4,136	.433799728	Root MSE	=	.5554

colgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsize	-.0561173	.0163799	-3.43	0.001	-.0882308	-.0240039
hsizesq	.0047609	.0022534	2.11	0.035	.0003431	.0091786
hsperc	-.0129015	.0005685	-22.69	0.000	-.014016	-.0117869
sat	.0016045	.0000661	24.27	0.000	.0014749	.0017341
female	.147631	.0179455	8.23	0.000	.1124481	.182814
_cons	1.286809	.0788179	16.33	0.000	1.132283	1.441335

图 5.

构建 F 统计量做联合假设检验：

$$F = \frac{(SSR_r - SSR_{ur})/(df_r - df_{ur})}{SSR_{ur}/df_{ur}} = \frac{(1274.29 - 1269.37)/2}{1269.37/(4137 - 8)}$$

可得 F 统计量的值为8.0，大于显著性水平为 1%的临界值，因此可以在显著水平为1% 的情况下拒绝原假设，也即：可以拒绝“女性运动员与女性非运动员的成绩没有差异”的原假设。