计量经济学第 2 次作业

Karry

1　多元线性回归

（a）根据所给结果，判断年龄（age）与教育水平（eduy）之间的相关性

答：

根据所给结果可以判断出年龄（age）与教育水平（eduy）之间的线性相关性并不高

原因如下：

多元回归中在MLR.1~MLR.5之下，以自变量的样本值为条件，均有：

$$Var(\hat{\beta}_i) = \frac{\sigma^2}{SST_j(1-R_j^2)}$$

由所给结果可以发现：工资水平只对教育水平回归时 $\hat{\beta}_{eduy}$ 的标准差(0.000788)与工资水平对年龄和教育水平同时回归时$\hat{\beta}_{eduy}$ 的标准差(0.000765)差别不大。可以推得 R_{eduy} 并不大也就是说年龄和教育书评之间的线性相关性并不高。

当然我们也对这一结论进行了验证，即将年龄（age）对教育水平（eduy）做回归得到下图1.中的结果：可以看到二者之间进行回归的 R^2 很小，即线性相关度很小。

```
reg eduy age
```

Source	SS	df	MS	Number of obs	=	31,237
				F(1, 31235)	=	1595.72
Model	11016.0209	1	11016.0209	Prob > F	=	0.0000
Residual	215630.588	31,235	6.90349248	R-squared	=	0.0486
				Adj R-squared	=	0.0486
Total	226646.609	31,236	7.25594214	Root MSE	=	2.6274

eduy	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0698232	.0017479	-39.95	0.000	-.0732492	-.0663972
_cons	14.61384	.0654539	223.27	0.000	14.48554	14.74213

图 1. 教育水平对年龄回归结果

（b） 使用数据emp2007.dta重复以上的回归结果，并说明age的系数含义。

答：

复现结果如图2. 图3.所示

```
. reg lnw eduy
```

Source	SS	df	MS	Number of obs	=	31,237
				F(1, 31235)	=	4498.68
Model	633.045232	1	633.045232	Prob > F	=	0.0000
Residual	4395.32671	31,235	.140717999	R-squared	=	0.1259
				Adj R-squared	=	0.1259
Total	5028.37194	31,236	.160980021	Root MSE	=	.37512

lnw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
eduy	.0528497	.000788	67.07	0.000	.0513053	.0543941
_cons	1.602484	.0097426	164.48	0.000	1.583388	1.62158

图2. 工资水平对数对教育水平进行回归的结果

```
. reg lnw eduy age
```

Source	SS	df	MS	Number of obs	=	31,237
				F(2, 31234)	=	4299.80
Model	1085.56489	2	542.782447	Prob > F	=	0.0000
Residual	3942.80705	31,234	.126234458	R-squared	=	0.2159
				Adj R-squared	=	0.2158
Total	5028.37194	31,236	.160980021	Root MSE	=	.35529

lnw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
eduy	.0629493	.0007651	82.27	0.000	.0614496	.0644489
age	.0145086	.0002423	59.87	0.000	.0140336	.0149836
_cons	.9515069	.0142606	66.72	0.000	.9235556	.9794582

图3. 工资水平对教育水平和年龄进行回归的结果

age 的系数含义：age 每增加一单位（一岁），工资提高 1.451%

（c）估计多元线性模型 $ln(w) = \beta_0 + \beta_1 eduy + \beta_2 age + \beta_3 age^2 + \mu$ 说明age量的边际效应是什么？分别计算当年龄等于20与50时，年龄的边际效应大小。

答：

多元线性模型的回归结果如下图 4. 所示：

```
. reg lnw eduy age age_squared
```

Source	SS	df	MS	Number of obs	=	31,237
				F(3, 31233)	=	2965.09
Model	1114.64492	3	371.548308	Prob > F	=	0.0000
Residual	3913.72702	31,233	.125307432	R-squared	=	0.2217
				Adj R-squared	=	0.2216
Total	5028.37194	31,236	.160980021	Root MSE	=	.35399

lnw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
eduy	.0618749	.0007656	80.82	0.000	.0603744	.0633754
age	.0473089	.0021666	21.84	0.000	.0430622	.0515555
age_squared	-.0004588	.0000301	-15.23	0.000	-.0005178	-.0003997
_cons	.4115976	.0381834	10.78	0.000	.3367567	.4864385

图 4. 多元线性回归结果

> age 量的边际效应就是对 lnw 对 age 求导，由回归结果可知
>
> $$边际效应 = \frac{\partial lnw}{\partial age} = \beta_2 + 2\beta_3 age = 0.0473 - 0.0004588 \times 2\ age$$
>
> 这个结果说明年龄（age）对工资水平具有递减影响
>
> - age = 20 时代入上式可求得 边际效应 = 0.0289
> - age = 50 时代入上式可求得 边际效应 = 0.00142

（d）使用两步法估计（c）中 β_1，将其与（c)中的结果比较。

> 答：
>
> 根据提示，我首先将 $eduy$ 对 age 以及 age^2 进行回归并计算回归残差 x_1 然后将 lnw 对残差 x_1 进行回归，得到如图 5. 所示的结果

```
. reg lnw x_1
```

Source	SS	df	MS	Number of obs	=	31,237
				F(1, 31235)	=	6073.16
Model	818.536765	1	818.536765	Prob > F	=	0.0000
Residual	4209.83518	31,235	.13477942	R-squared	=	0.1628
				Adj R-squared	=	0.1628
Total	5028.37194	31,236	.160980021	Root MSE	=	.36712

lnw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x_1	.0618749	.000794	77.93	0.000	.0603187	.0634311
_cons	2.240249	.0020772	1078.50	0.000	2.236178	2.244321

图5. 两步法估计

> 与（c）中结果相比较可以发现
>
> - β_1 的估计值完全相同
> - 但是 β_1 的方差有所差别，我们目前的猜测是因为本次回归较比（c）中回归忽略了残差。

（e）构建一个新的变量 $exp = age - eduy - 6$，来表示工作经验（你的工作经验大约等于年龄去掉上学的年限和学前年限）。估计多元线性模型

$$ln(w) = \beta_0 + \beta_1 eduy + \beta_2 age + \beta_3 age^2 + \beta_4 exp + \mu$$

报告其结果，发生了什么异常？请做出解释。

> 答：
>
> 我们在图 7. 中报告了该多元线性模型的回归结果

```
. reg lnw eduy age age_squred exp
note: age omitted because of collinearity
```

Source	SS	df	MS	Number of obs	=	31,237
				F(3, 31233)	=	2965.09
Model	1114.64492	3	371.548308	Prob > F	=	0.0000
Residual	3913.72702	31,233	.125307432	R-squared	=	0.2217
				Adj R-squared	=	0.2216
Total	5028.37194	31,236	.160980021	Root MSE	=	.35399

lnw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
eduy	.1091838	.0022489	48.55	0.000	.1047757	.1135918
age	0	(omitted)				
age_squred	-.0004588	.0000301	-15.23	0.000	-.0005178	-.0003997
exp	.0473089	.0021666	21.84	0.000	.0430622	.0515555
_cons	.6954508	.0260741	26.67	0.000	.6443445	.7465571

图 7. 加入工作经验后的回归结果

在图 7. 所报告的回归结果中我们发现：age 的系数估计量为0，标准差为无穷大。

解释如下：构建新变量 $exp = age - eduy - 6$ 直接导致了高斯马尔可夫假设中的 MLR.3即不存在完全共线性假设无法满足，因为自变量 exp，age，$eduy$ 之间存在严格的线性关系，模型遇到了完全共线性的问题，不能由 OLS 来估计。

age 的系数估计量为 0 这一点可以在其他条件不变的模式下解释。正常情况下 age的系数表示在其他条件（eduy，exp）不变的情况下，age 的增加对工工资水平的影响。但是 eduy,exp 不变的话 age 根本就不会改变，所以毫无意义。

age 标准差为无穷大这一点，可以由 $Var(\hat{\beta}_i) = \frac{\sigma^2}{SST_j(1-R_j^2)}$ 来解释，因为 $R_{age} = 1$

2 高管 CEO 薪水

考虑企业高管CEO的薪水（salary）与企业销售收入、股权回报率（return on equity，roe）以及企业股票的收益率（return on firm's stock，ros）之间的关系，建立以下的多元线性方程：

$$ln(salary) = \beta_0 + \beta_1 ln(sales) + \beta_2 roe + \beta_3 ros + \mu$$

（a）基于模型参数，建立原假设H0：在控制sales和roe的条件下，ros 对CEO的薪水没有关系；备选假设H1为：企业股票表现越好，企业CEO的薪水越高。

（b）基于以上模型，OLS估计结果如下（括号里面是标准差）

$$\log(salary) = 4.32 + 2.80\log(sales) + 0.174roe + 0.00024ros$$

（0.32）（0.035）　　（0.0041）　（0.00054）

n=209，R^2=0.283。

如果ros增加 50，CEO的薪水增加多少？你认为ros对高管薪金的影响大吗？

> 答：在其他条件不变时 ros 增加 50， CEO 的薪水增加1.2 %
>
> 我认为 ros 对高管薪金的影响相对不大，因为 ros 的取值为[0, 100]也就是说 ros 最大能增加 100，此时在其他条件不变的情况下，CEO 薪水增加不过 2.4%。当然如果一个 CEO 年入千万的话，这一部分也将近几十万了，但是仍然占比依然很小。

（c）检验原假设：ros对薪金没有影响，其备择假设为：ros对薪金有正的影响。10%的显著水平下，临界值为 1.282.

> 答：
>
> 由回归结果可知： $t(ros) = \frac{0.00024}{0.00054} = 0.444 < 1.282$
>
> 也就是说在为 10% 的显著性水下，我们不能拒绝原假设。

（d）基于以上的分析，你认为ros是解释CES薪金的重要因素？是否应该将其从模型中删除？

> 答：由（b）（c）中的分析我认为 ros 并不是解释 CES 薪金的重要因素。
>
> - ros 的经济显著性并不高，这一点体现在回归结果中 ros 的系数的绝对大小，这一点在 （c） 中有所说明
> - ros 统计学上并不显著。
>
> 但是我们不能将其从模型中剔除。因为统计学上显著与否并不是我们判断是否要加入变量的标准，因为不能拒绝原假设也并不意味着可以接受，同时我们怀疑roe和ros之间存在一定的线性相关性导致了不显著的结果。但是：我们重点关注的还是 ros 前面的系数所带来的因果关系。

3　房价预测

基于数据hprice1.dta，我们将通过考察实际房价水平与评估（预期）房价之间的关系来检验房价预期是否理性。具体模型如下：

$$Price = \beta_0 + \beta_1 assess + u$$

（a）基于OLS估计该模型。如果预期是理性的，那么$\beta_0 = 0$ 且 $\beta_1 = 1$。接下来首先检验假设H0：$\beta_0 = 0$（备选假设$\beta \neq 0$）；然后检验H1：$\beta_1 = 1$（备选假设$\beta_1 \neq 1$)。以上检验的显著性水平要求均为 5% 。你的结论是什么？

答：

基于 OLS 估计该模型的结果如图8. 所示

```
. reg price assess
```

Source	SS	df	MS	Number of obs	=	88
				F(1, 86)	=	390.54
Model	752209.994	1	752209.994	Prob > F	=	0.0000
Residual	165644.511	86	1926.09897	R-squared	=	0.8195
				Adj R-squared	=	0.8174
Total	917854.506	87	10550.0518	Root MSE	=	43.887

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
assess	.9755538	.0493652	19.76	0.000	.8774191	1.073689
_cons	-14.47179	16.27339	-0.89	0.376	-46.82221	17.87863

图8.　基于 OLS 估计模型的结果

注意到本模型的自由度为 88 - 2 =86 由于此自由度已经很大，足以使用标准正态作为近似，所以显著性水平为 5% 的临界值约为 1.96。

首先检验假设H0：$\beta_0 = 0$

可以由回归结果中看到：$\hat{\beta}_0 = -14.47 \quad se(\hat{\beta}_0) = 16.27 \ t = -0.89 > -1.96$

因此：在 5% 的显著性水平上无法拒绝 H0，即在显著性水平为 5% 时截距项并不显著异于 0

然后检验假设H1：$\beta_1 = 1$

不同于H0，此时的 $t = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} = \frac{0.98-1}{0.049} = -0.41 > -1.96$

因此：在 5% 的显著水平上无法拒绝 H1，即在显著性水平为 5% 时β_1 并不显著异于 1

> 综上：我们可以初步判断预期是较为理性的。

（b）对联合假设H0：$\beta_0 = 0$ 且$\beta_1 = 1$ 进行 F 统计检验。你是否可以在5%的显著性水平下拒绝原假设？如果是在1%的显著性水平下呢？（提示：需要通过计算不受约束与受约束条件下的均方和来获得统计值)

> 答：
>
> 题中给出的原模型为不受约束模型即 $Price = \beta_0 + \beta_1 assess + u$
>
> 注意到联合假设是$\beta_0 = 0$ 且 $\beta_1 = 1$
>
> 因此针对假设检验设定得受约束模型为：$Price = assess + u$ ，现在为了施加 $\beta_1 = 1$的约束我们选择估计如下模型：
>
> $$Price - asess = u$$
>
> 构建 F 统计量
>
> $$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$
>
> 其中：
>
> - SSR_r是受约束模型的残差平方和，SSR_{ur} 是不受约束的残差平方和
> - 分子自由度 $q = 2$，分母自由度 $n - k - 1 = 86$

```
. gen tmp=price-assess

. reg tmp
```

Source	SS	df	MS	Number of obs	=	88
				F(0, 87)	=	0.00
Model	0	0	.	Prob > F	=	.
Residual	166116.855	87	1909.38913	R-squared	=	0.0000
				Adj R-squared	=	0.0000
Total	166116.855	87	1909.38913	Root MSE	=	43.697

tmp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons	-22.19033	4.658069	-4.76	0.000	-31.44874	-12.93191

图 9. 受约束模型的回归结果

因此可得 F 统计量为：

$$F = \frac{(166116.86 - 165644.51)/2}{165644.51/86} = 0.123$$

而对于分子自由度为 2 分母自由度为 86

- 显著性水平为 5% 的临界值为 3.1
- 显著性水平为 1% 的临界值为 4.86 （均由 Excel 中的 FINV 函数求得）

可以看出显著性水平不论是 1% 还是 5% F 统计值都远小于临界值，故不能在 5% 或 1% 的显著性水平下拒绝原假设

（c）接下来估计模型（sqrft代表房间的总面积，lotsize代表房屋地皮的尺寸大小，bdrms代表卧室的数目）：

$$Price = \beta_0 + \beta_1 assess + \beta_2 sqrft + \beta_3 lotsize + \beta_4 bdrms + u$$

计算基于 R^2 的 F 统计量来检验联合假设 H0：$\beta_2 = 0, \beta_3 = 0, \beta_4 = 0$。P值为多少？

答：本题思路和（b）题完全相同，只不过计算 F 统计量时是基于 R^2 来算，注意：本题计算 F 时分子的自由度为 3 ，分母的自由度为 83

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)}$$

此时不受约束模型变为了

$Price = \beta_0 + \beta_1 assess + \beta_2 sqrft + \beta_3 lotsize + \beta_4 bdrms + u$

而题中给出的 $Price = \beta_0 + \beta_1 assess + u$ 变成了受约束模型

基于 OLS 估计本不受约束模型的结果如图 10.

```
. reg price assess sqrft lotsize bdrms
```

Source	SS	df	MS	Number of obs	=	88
				F(4, 83)	=	100.74
Model	761089.801	4	190272.45	Prob > F	=	0.0000
Residual	156764.704	83	1888.73138	R-squared	=	0.8292
				Adj R-squared	=	0.8210
Total	917854.506	87	10550.0518	Root MSE	=	43.46

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
assess	.9082991	.1040386	8.73	0.000	.7013706	1.115228
sqrft	-.0005175	.0170849	-0.03	0.976	-.0344986	.0334636
lotsize	.0005867	.0004963	1.18	0.240	-.0004004	.0015738
bdrms	11.60249	6.549515	1.77	0.080	-1.424233	24.62921
_cons	-38.88702	21.49853	-1.81	0.074	-81.64673	3.872696

图10. 不受约束模型估计结果

> 可以带入公式计算 F = 1.57，结合分子的自由度为 3 ，分母的自由度为 83 可得
>
> P = 0.203 （由 Excel 中的 FDIST 求得）

（d）如果变量 price 的方差会随着 sqrft、lotsize 或者 bdrms 的变化而变化，是否会影响 F 检验的有效性？

> 答：会影响 F 检验的有效性。
>
> 如果被解释变量 Price 的方差会随着解释变量 sqrft、lotsize、bdrms 的变化而变化，那么也就意味着 $D(u|x)$ 不再是一个常数，本线性回归模型就不再满足同方差性。这样的话，F 统计量便无法成为 F 统计量（因为在公式推导过程中有一步骤需要分子分母同时约去 $D(u|x)$），F 检验也就不再有效了。

（e）估计以下的线性方程：

$$lnPrice = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

如果增加一个150-square-root的卧室，房价变动的百分比是多少？（提示：增加一个卧室，同时也会增加房间的整体面积）

> 答：基于 OLS 对本方程的估计结果如图 11.
>
> ```
> . reg lprice sqrft bdrms
> ```
>
Source	SS	df	MS	Number of obs	=	88
				F(2, 85)	=	60.73
Model	4.71671468	2	2.35835734	Prob > F	=	0.0000
Residual	3.30088884	85	.038833986	R-squared	=	0.5883
				Adj R-squared	=	0.5786
Total	8.01760352	87	.092156362	Root MSE	=	.19706

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sqrft	.0003794	.0000432	8.78	0.000	.0002935	.0004654
bdrms	.0288844	.0296433	0.97	0.333	-.0300543	.0878232
_cons	4.766027	.0970445	49.11	0.000	4.573077	4.958978
>
> 图11. 对该线性方程的估计结果
>
> 可以看出在其他条件不变的情况下，如果增加一个150-square-root的卧室也就意味着 bdrms 增加 1 同时 sqrft 增加 150，则房价变动的百分比应该为：
>
> $$0.0003794 \times 150 + 0.0288844 \times 1 = 8.58\%$$

（f）通过改写（e）中的模型，以便于你可以直接检验“增加一个150-square-root的卧室对房价的影响大小”。构造估计值95%的置信区间。

答：设我们要估计的为 θ 则 $\theta = 150\beta_1 + \beta_2$ 那么我们用 θ 和 β_1 来表示 β_2 的话就有 $\beta_2 = \theta - 150\beta_1$ 将此代入(e)中的方程可得：

$$lnPrice = \beta_0 + \beta_1 sqrft + (\theta - 150\beta_1)bdrms + u$$

进而：

$$lnPrice = \beta_0 + \beta_1(sqrft - 150bdrms) + \theta bdrms + u$$

上式就是我们对（e）中的模型改写的结果，可以看到这其中的 θ 就表示其他条件不变时增加一个150-square-root的卧室对房价的影响大小。我们在图12.中报告了对该模型的估计结果，可以看出在此估计出的 $\hat{\theta}$ 和上一题中的结果完全相同。

```
. gen tmp= sqrft-150*bdrms

. reg lprice tmp bdrms
```

Source	SS	df	MS	Number of obs	=	88
				F(2, 85)	=	60.73
Model	4.71671468	2	2.35835734	Prob > F	=	0.0000
Residual	3.30088884	85	.038833986	R-squared	=	0.5883
				Adj R-squared	=	0.5786
Total	8.01760352	87	.092156362	Root MSE	=	.19706

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tmp	.0003794	.0000432	8.78	0.000	.0002935	.0004654
bdrms	.0858013	.0267675	3.21	0.002	.0325804	.1390223
_cons	4.766027	.0970445	49.11	0.000	4.573077	4.958978

图12. 该模型估计结果

由图12. 的报告我们可以直接得到 $\hat{\theta}$ 95% 的置信区间为 [0.0326, 0.1390]

当然我们也可以由置信区间的定义算得 $\hat{\theta}$ 95%的置信区间为

$$[\hat{\theta} \pm c * se(\hat{\theta})] = [0.0326, 0.1390]$$