《计量经济学：Stata 实验专题》期中测试

命题人：赵震宇　审题人：攸频、葛家和

测试时长：120 分钟　满分：100 分

测试说明：

1. 本次测试共五道大题，其中前四道为必做题，每题 25 分，共计 100 分，每道大题下共有 5-10 个小问，每小问 5/2.5 分，视 Stata 实现方法优劣给予不同评分，过程、结果错误不得分；第五道为选做题，两小题各 10 分共 20 分，计入总分但总分不超过 100 分。
2. 最后只提交 do 文件、描述性统计和回归结果的 word 文件，打包成一个压缩包并命名为学号+姓名如“2310000+张三”提交；do 文档里包括所有的代码，Stata 运行结果、问题答案阐述以注释形式呈现，注意标明题号。
3. 建议采用“adopath +”命令将文件路径设置到给大家发的 adofiles 下，否则可能需要安装一部分外部命令。
4. 测试形式为开卷，可以使用必要的纸质资料和电子资料，可以联网搜索，因为信息检索与甄别能力同样非常重要，但请分配好答题时间。
5. 禁止使用手机、私人电脑等通讯设备，请独立完成，不得交流讨论、分工合作。
6. 不反对使用 AI，但请注意 AI 给出的结果并不完全是正确的，让 AI 辅助你思考而非替代你思考！锻炼与 AI 对话的能力！

一、数据清洗

1. （数据导入）在国泰安上下载了 2023 年上市公司全 A 股（剔除 ST）的三大财务报表，导入资产负债表（“FS_Combas.xlsx”文件）到 Stata 中。（Hint：观察文件前三行分别是变量的英文名、变量的中文名、单位，你将如何处理这三行数据？）

2. （数据类型转换）将股票代码 Stkcd 转换为数值型变量，并生成新的变量 id；将统计截止日期 Accper 转换成数值型“31dec2023”或“20231231”格式。

3. （变量生成与清洗）生成一个新变量资产负债率 lev；在数据清洗过程中，需要对异常值和极端值进行截尾或缩尾处理，判断 lev 是否存在异常值，你将如何处理呢？（Hint：资产负债率=总负债/总资产，资产负债率通常是一个 0-100%之间的数）

4. （字符提取与删除）在利润表（“FS_Comins.dta”文件）中，不小心混入了一部分特别处理（Special Treatment）的上市公司，它们的证券简称 ShortName 都以 ST 或*ST 开头，剔除这一部分上市公司，保存数据并替换原文件。

5. （数据追加与合并）现金流量表（“FS_Comscfd.dta”文件）被拆分成了两部分数据，首先将“FS_Comscfd_add.dta“文件中的上市公司追加到“FS_Comscfd.dta“文件中，保

存数据并替换原文件；然后以资产负债表为主数据，合并这三张财务报表，只保留观测值仅在主数据中存在的上市公司。（Hint：根据股票代码 Stkcd 作为唯一标识）

二、数据统计

为避免你在上一道题目中得到的合并后的数据集有误影响接下来的分析，本题换用 Stata 自带的 auto.dta 数据，首先导入该数据。

1. rep78 存在缺失值吗？有重复值的 mpg 的值有几个？解释使用 duplicates list、duplicates report 和 duplicates example 时的不同报告结果的含义。
2. 对 headroom 进行中心化，根据汽车是否是国产（foreign）分组对 trunk 进行标准化，对 turn 进行归一化。（Hint：归一化=(变量 x-最小值)/(最大值-最小值)）
3. 如果将 price 作为被解释变量，mpg weight length gear_ratio 作为解释变量进行回归，检验模型的异方差性、检验 weight 和 length 的联合显著性；检验 displacement 的组间均值差异（同样按照汽车是否是国产分组）。
4. 计算 weight 和 length 的 Pearson 相关系数，上一问回归模型中存在严重的多重共线性问题吗？你将如何处理呢？
5. 将 price mpg weight length foreign 的描述性统计结果输出并保存到一个 word 文件中（命名为“output1.docx”），描述性统计通常包含观测值、平均值、标准差、最小值和最大值等，除观测值外其它统计值都保留两位小数。

三、蒙特卡罗模拟

1. 清除之前的数据集，设置种子数 10101，样本量为 500；

2. 生成解释变量教育年限 educ（单位：年），educ 服从均值为 12，标准差为 3 的正态分布，使用 round ()函数将 educ 四舍五入到最接近的整数；

3. 生成控制变量工作经验 exper（单位：年）（与 educ 负相关）：exper = 20 – educ + u，u~N(0,2)，画出 exper 的直方图；

4. 生成能力 ability（理论上不可观测，与 educ 相关）：ability = 1.5*log(educ) + u_1，u_1~N(0,1)；取对数有什么好处？

5. 生成每天的工作时间 hours（单位：小时）服从[0,10]的均匀分布；生成一个用于判断是否符合八小时工作制的虚拟变量 is8h，如果工作时间在 8 小时以内（含 8 小时）则 is8h = 1，否则= 0；

6. 生成被解释变量年薪 wage（单位：万元），wage = 15 + 0.8*educ + 1.2*exper + 2*ability + e，e~ N(0,3) ；简单回归模型（记为 m1）中被解释变量为 wage，解释变量为 educ；

7. 加入 exper 到回归方程中，记多元回归模型为 m2；在一张表里同时列示出两个模型，如果遗漏掉 exper 偏误如何？

8. 在模型 m2 中，求出残差 uhat 并检验其正态性；

9. 画出 uhat 与 ability 的散点图和拟合线，你能得出什么结论？

10. 预测出一个大学毕业(educ = 16)、有一定实习经验(exper = 2)的学生的工资。

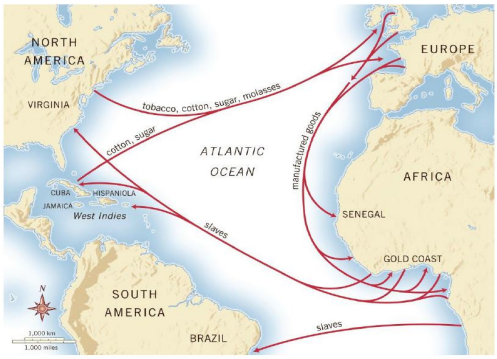
四、回归分析

我们来尝试复现一篇 2011 年发表在经济学顶刊 AER 上的经典的截面数据文章的部分内容，标题是《奴隶贸易和非洲不信任的起源》。

发展经济学里有两大关注的问题：一是部分国家和地区为何落后，二是后进经济体如何发展实现赶超。

现代非洲国家的普遍不发达源于历史的奴隶贸易，即一个国家历史上出口奴隶的数量与该国当今的经济表现存在强烈的负相关关系（Nunn, 2008 QJE），这篇文章在此基础上探讨了具体的影响机制。

良好的信任能促进合作并推动创新与经济发展。文章指出历史上的奴隶贸易引起了非洲内部的群体不信任并持续至今，这是一个恶循环：信任→受到损失→不信任→进一步不信任→世世代代不信任→基因里就认为不信任是正确的。



三角贸易示意图

回归方程和变量说明如下：

$$trust_{i,e,d,c} = \alpha_c + \beta slave\ exports_e + \mathbf{X}'_{i,e,d,c}\mathbf{\Gamma} + \mathbf{X}'_{d,c}\mathbf{\Omega} + \mathbf{X}'_e\mathbf{\Phi} + \varepsilon_{i,e,d,c},$$

	变量	说明
下标	i	个体（individual）
	e	种族（ethnic）
	d	地区（district）
	c	国家（country）
被解释变量	$trust_{i,e,d,c}$	共 5 种衡量信任的方式
核心解释变量	$slave\ exports_e$	共 6 种衡量奴隶出口的方式
截距项	α_c	国家固定效应（如政府法规影响）

控制变量	$\mathbf{X}'_{i,e,d,c}$	**个体层面的协变量** （年龄、性别、城乡、生活条件、学历、宗教、职业）
	$\mathbf{X}'_{d}$	**地区层面的协变量** （种族构成、人口比例）
	$\mathbf{X}'_{e}$	**种族层面的协变量** （历史特征、殖民统治的影响）
误差项	$\varepsilon_{i,e,d,c}$	

注：由于衡量标准在地区和种族层面有所不同，$\mathbf{X}'_{d}$ 和 $\mathbf{X}'_{d,c}$ 符号表示有出入。

本题使用“Trust_OLS.dta”数据，补全“Trust_OLS.do”中的代码，具体要求见最后一页。

五、附加题

（一）正则表达式初探

正则表达式（regular expression）描述了一种字符串匹配的模式，可以用来检查一个串是否含有某种子串、将匹配的子串替换或者从某个串中取出符合某个条件的子串等，即根据规定的筛选方式从文本当中挑选符合要求的字段，常用符号如下：

符号	含义
\	转义符，对于一些特殊的字符需要进行转义，不然无法匹配，例如括号等。
^	是否需要从待匹配字符串开始位置进行匹配，即只取头部。
$	是否需要以待匹配字符串结尾位置作为匹配结尾，即只取尾部。
*	对前面表达式匹配零次或多次，相当于{0,}
+	对前面表达式匹配一次或多次，相当于{1,}
?	对前面表达式匹配零次或一次。
{n}	对前面表达式匹配 n 次。
{n,}	对前面表达式至少匹配 n 次。
{m,n}	对前面表达式至少匹配 m 次，至多匹配 n 次。
|	表示或，A|B 表示匹配 A 或者 B。
[0-9]	匹配数字，相当于\d。
[^0-9]	匹配非数字，相当于\D。
[a-z]	匹配小写字母。
[A-Z]	匹配大写字母。
[a-zA-Z0-9]	匹配字母与数字。

在 Stata 中与正则表达相关的常用函数是 ustrregexm 与 ustrregexs，前者主要用于正则表达式的匹配，后者主要用于对匹配结果的截取，按照特定正则表达规则提取文本变量时，往往需要两者配合使用。

观察“regular_expression.do”中给出的例子，完成后面问题的作答。

（二）惩罚回归（岭回归、Lasso 回归、弹性网回归）[①]

9.1 使用 UCI Machine Learning Repository 的葡萄牙高中数学成绩数据 student-mat.csv，进行惩罚回归，其中，响应变量（即被解释变量）G3 为期末成绩，而特征变量（即解释变量）包含学生、学校以及父母的一系列变量。

(1)由于此 CSV 文件以分号";"分割，故使用命令"import delimited "student-mat.csv", delimiter(";") clear"载入数据，并考察此数据框的形状与前 5 个观测值；

(2)从数据集中去掉变量 G1 与 G2，因为这是同一学期的前两阶段成绩，且与 G3 高度相关；

(3)画响应变量 G3 的直方图；

(4)将数据矩阵中的分类变量都变为虚拟变量；

(5)将所有特征变量标准化；

(6)考虑惩罚参数 λ[②]，画岭回归的系数路径（Stata 命令 lasso2[③]）；

(7)通过 10 折交叉验证[④]（Stata 命令 cvlasso，设置种子数 123），使用最优 λ[⑤]进行岭回归；

(8)画 Lasso 回归的系数路径；

(9)通过 10 折交叉验证，使用最优 λ 进行 Lasso 回归；

(10)考虑调节参数 α 的取值[0，0.001，0.01，0.1，0.5，1]，对 λ 也进行交叉验证（设置种子数 123），确定最优的 λ 与 α。

---

① 本题改编自陈强《机器学习及 Python 应用》第 9 章习题

② Stata 中用 lambda 表示正则力度

③ Stata 中用 alpha 表示 L1、L2 权重，岭回归：alpha(0)、Lasso 回归：alpha(1)、弹性网：如 alpha(0.5)

④ K 折交叉验证（K-fold cross validation），默认 K=10

⑤ Stata 选项 lopt：选择使均方预测误差（Mean Squared Prediction Error，MSPE）最小的 λ

```
*- 本页对应第四大题，我们来复现原文第III章的两张表格（见"回归结果.pdf"）

*- 被解释变量、解释变量（见"描述性统计.pdf"）和控制变量如下：
    *- 被解释变量（共5种衡量信任的方式）包括：
        *- 对亲属的信任程度（trust_relatives）
        *- 对邻居的信任程度（trust_neighbors）
        *- 对地方政府的信任程度（trust_local_council）
        *- 对本族裔的信任程度（trust_intra_group）
        *- 对外族裔的信任程度（trust_inter_group）
    *- 解释变量（共6种衡量奴隶出口的方式）包括：
        *- 被掳走的奴隶数量（exports）
        *- 被掳走的人数/居住面积（export_area）
        *- 被掳走的人数/历史人口（export_pop）
        *- 将前三个解释变量分别+1取对数作为后三个解释变量
    *- 控制变量包括：
*- 个体层面的协变量包括age age^2 male urban living_conditions education religion occupation
*- 地区层面的协变量包括种族构成（district_ethnic_frac）和人口比例（frac_ethnicity_in_district）
*- 种族层面的协变量暂时用不到

Question1：生成后三个解释变量（不要求使用循环语句）
Question2：使用全局暂元存储被解释变量（depend_var）、解释变量（independ_var）和基础控制变量
    *- 基础控制变量（baseline_controls）包括个体和地区层面的协变量加上国家代码（isocode）
    *- 要求使用c.age构造一次项和二次项
    *- living_conditions、education、religion、occupation、isocode需要设置成虚拟变量

*********************************
*** Table 1: Trust Neighbors ***
*********************************
*- 被解释变量使用trust_neighbors
*- 解释变量将6种衡量奴隶出口的方式都纳入进来
*- 加入基础控制变量（baseline_controls）

*- 方括号里的标准误是聚类到一维种族（murdock）层面
*- 圆括号里的标准误是聚类到二维种族和地区层面
*- 花括号里的标准误报告的是二维空间自相关调整的标准误

Question3：导出一个名为output2的word文件（outreg2可替换为其它命令）
    *- 为简化分析，输出结果里包含正确的系数和第一个（方括号里的）标准误即可
    *- Hint：聚类调整的标准误需要在回归后加入cluster()，括号内加入聚类层面

***********************************
*** Table 2: All Trust Measures ***
***********************************
*- 被解释变量将5种衡量信任的方式都纳入进来
*- 解释变量采用ln(1+export/area)
*- 加入基础控制变量，依然聚类到种族层面

*- 不同于表1，表2我们使用循环语句的嵌套实现
*- 同时报告含"Yes"的三行和N、r2

Question4：补全第一个foreach、reg、后两个estadd、eststo、esttab后的代码
    *- Hint：第一层循环加入被解释变量、仿照第一处estadd，最后的结果be like:

/*  ------------------------------------------------------------------------------------
                        (1)             (2)             (3)             (4)             (5)
                trust_rela~s    trust_neig~s    trust_loca~l    trust_intr~p    trust_inte~p
    ------------------------------------------------------------------------------------
    ln_export_~a       -0.133***       -0.159***       -0.111***       -0.144***       -0.097***
                      (0.036)         (0.034)         (0.022)         (0.031)         (0.028)
    ------------------------------------------------------------------------------------
    Individual~s          Yes             Yes             Yes             Yes             Yes
    District_c~s          Yes             Yes             Yes             Yes             Yes
    Country_fi~s          Yes             Yes             Yes             Yes             Yes
    N                   20062           20027           19733           19952           19765
    r2                   0.13            0.16            0.20            0.14            0.11
    ------------------------------------------------------------------------------------
    Standard errors in parentheses
    * p<0.05, ** p<0.01, *** p<0.001                                                    */
```