Mid-term Review

Some useful Stata command

1. 数据处理

- 打开文件：`use 文件路径\文件名.dta, clear`
- 变量标签：`label variable x xlabel`
- 审视数据：`describe`，`list`
- 统计特征：`summarize`，`tabulate`，`correlate`，`pwcorr`，`spearman`
- 画图：`histogram x, width(1) frequency`，`kdensity x`，`scatter x1 x2`，`twoway (scatter x1 x2) (lfit x1 x2)`，`twoway (scatter x1 x2) (qfit x1 x2)`
- 生成新变量：`gen lnx = log(x)`，`gen xsq =  x^2`，`gen x = .`，`replace x = 1 if x > 1`，`rename x x1`，`drop ln*`
- 计算器功能：`display log(2)`

取对数！取对数？

- 缩小数据之间的绝对差异；避免个别极端值的影响
- 尽可能满足经典线性模型假定（Classic Linear Model）
  - 避免共线性
  - 避免异方差
  - 尽可能符合正态分布
- 经济学意义
  - 用线性模型估计非线性关系
  - 直接估计弹性/半弹性
  - 取对数背后的经济理论模型
    - CD 生产函数
    - 某些变量本身就以对数形式存在

2. 线性回归分析

- `regress y $x`
- `regress y $x, noc`
- `regress y $x if q >= 6000`
- `predict yhat`
- `predict uhat, residual`
- `display 1/_b[x1]`

- `test x1 = 1`
- `test (x1 = 1) (x2 + x3 + x4 = 1)`
- `test x1 x2`
- `testnl _b[x1] = _b[x2]^2`

Stata：标准误！标准误！

- `reg y $x, robust`
- `reg y $x, cluster(id)`
- `reg y $x, vce(robust)`
- `reg y $x, vce(cluster id)`
- `reg y $x, vce(bootstrap)`

Stata：聚类标准误代码介绍：

聚类调整标准误 (cluster)的基本思想是放宽了随机误差项独立同分布的假定，允许组内个体的干扰项之间存在相关性，但不同组个体的干扰项之间彼此不相关。

当处理的分配机制或抽样过程存在聚类性，则需要在该层面对标准误进行聚类。例如，某一政策的实施与否在地级市层面的决策，且抽样过程也是以地级市为单位 (尽管研究的个体单位可能是地级市中的企业)，那么标准误就应该聚类到地级市层面。

一维聚类标准误

```
*截面数据，在个体层面进行聚类，以下两种写法等价
reg y x, cluster(id)
reg y x, vce(cluster id)
```

`cluster(id)`, `cluster(industry)`, `cluster(area)`

二维（双向）/多维聚类标准误

```
reg y $x, vce(cluster area industry)
```

```
cgmreg y $x, cluster(id year)
vce2way reg y $x, cluster(id year)
vcemway reg y $x, cluster(id year)
```

3. 正态分布检验

- `hist x, normal`
- `kdensity x, normal`
- `qnorm x`
- `jb6 x`

```
sum x, detail
di (r(N)/6)*(r(skewness)*2)+[(1/4)*(r(kurtosis)-3)^2]
di chi2tail()
```

- `sktest x`
- `swilk x`
- `sfrancia x`

4. 异方差处理

- 画残差图：`rvfplot`
- 怀特检验：`estat imtest, white`，`whitetst`
- BP检验：
  - `estat hettest`
  - `estat hettest, rhs`
  - `estat hettest [varlist]`
  - `estat hettest, iid`
  - `estat hettest, rhs iid`
  - `estat hettest [varlist], iid`
- WLS：`reg y $x [aw = 1/var]`

5. 模型设定

- 遗漏变量：`estat ovtest`，`estat ovtest, rhs`
- 多重共线性：`estat vif`
- 极端数据：

```
predict lev, leverage
gsort -lev
sum lev
list lev in 1/3
```

Stata：离群值！离群值？离群值！

- 对数转换
- 缩尾

```
winsor x, gen(x_w) p(0.025)
winsor x, gen(x_wh) p(0.025) highonly
winsor x, gen(x_wl) p(0.025) lowonly
winsor2 x, replace cuts(2.5 97.5)
```

- 截尾

```
drop x if x >= 1
winsor2 x, cuts(2.5 97.5) trim
winsor2 x, cuts(2.5 100) trim
winsor2 x, cuts(0 97.5) trim
```

- 插值

6. 虚拟变量

- `gen d = (year >= 1978)`
- `tabulate province, gen(pr)`

```
xi: reg y x i.Dummy //Dummy是字符型
```

```
encode Dummy, gen(dummy)
reg y x i.dummy
```

7. 描述性统计和回归结果导出：

- `logout`
- `sum2docx`
- `asdoc`
- `esttab`
- `outreg2`
- `reg2docx`