

附录 1: stata 程序之互联网变量生成

***** 互联网文件生成*****

//B066 电子邮箱、B067 网址

clear all

save 互联网.dta, replace emptyok

local a : dir . files "a_*.dta"

local 互联网 B066 B067

foreach i of local a{

 use `i', clear

 foreach j of local 互联网 {

 cap gen `j' = ""

 }

 keep id FRDM B02 B07 YEAR `互联网'

 append using 互联网.dta

 save 互联网.dta, replace

}

merge m:1 id using 面板数据 id 提取.dta,keep(1 3) generate(面板 id)

//全角转半角

foreach i of local 互联网 {

 gen 备份`i' = `i'

 replace `i' = ustrlower(ustrtrim(`i'))

 foreach j of numlist 33/64 91/126 {

 replace `i' = usubinstr(`i',uchar(`j'+65248),uchar(`j'),.) //全角和半角相差 65248

 }

 replace `i' = usubinstr(`i',"—","-,.) //这两个不是全角和半角的区别

 replace `i' = ustrregexra(`i',"©|@|(a)|[a]","@")

 replace `i' = ustrregexra(`i',"@+","@")

 replace `i' = "" if `i' == "null"

}

sort id YEAR

save 互联网.dta, replace

*****一、B066、B067 内容互换

use 互联网.dta, clear

foreach i in B066 B067 {

```

cap drop 修复前`i'
gen 修复前`i' = `i',after(`i')
}

replace B067 = "" if B067 == B066 & ustrpos(B066,"@")
replace B066 = "" if B067 == B066 & ustrpos(B066,"@") == 0 & ustrregexm(B066,"www|http")
*****针对两个指标的观测值被合并为一个指标的观测值情况 (应该在 B066、B067 内容互换之前完成)

local                                regex                                =
"^((http.+|www.+)[[:punct:]]\s)+(com|net|edu|gov|org|biz|cn|hk|tw|mo|jp|kr|uk|de|fr|it|es|pt|nl|ru|us|ca|cc| 中
国)[[:punct:]]\s*([a-z0-9]([^\@\s]*?[a-z0-9])?\@\@.+)$"
gen 临时 = B066 if ustrregexm(B066,"`regex'") & B067 == ""
replace B067 = ustrregexs(1) + "." + ustrregexs(2) if ustrregexm(临时,"`regex'")
replace B066 = ustrregexs(3) if ustrregexm(临时,"`regex'")
//

local                                regex                                =
"^([a-z0-9]([^\@\s]*?[a-z0-9])?\@\@.+)(com|net|edu|gov|org|biz|cn|hk|tw|mo|jp|kr|uk|de|fr|it|es|pt|nl|ru|us|c
a|cc|中国)[[:punct:]]\s*((http.+|www.+)$"
cap drop 临时
gen 临时 = B066 if ustrregexm(B066,"`regex'") & B067 == ""
replace B067 = ustrregexs(4) if ustrregexm(临时,"`regex'")
replace B066 = ustrregexs(1) + ustrregexs(3) if ustrregexm(临时,"`regex'")
*****针对一个指标的观测值被分拆为两个指标的观测值情况 (应该在 B066、B067 内容互换之前完成)

cap drop 临时
gen 临时 = ustrright(B066,ustrlen(B066)-1) + "." + ustrleft(B067,ustrlen(B067)-1) if ustrleft(B066,1)
== `""' & ustrright(B067,1) == `""'

replace      临      时      =      B066      +      B067      if
ustrregexm(B066+B067,"[a-z0-9\-\_]{4,}?[[:punct:]]\s*(com|net|edu|gov|biz)$") &
ustrpos(B066,ustrright(B066+B067,3)) == 0 & ustrlen(B067) <= 6 & missing(B066,B067) == 0

replace      临      时      =      B066      +      B067      if
ustrregexm(B066+B067,"[a-z0-9\-\_]{4,}?[[:punct:]]\s*(cn|hk|tw|mo|jp|uk| 中 国 )$") &
ustrpos(B066,ustrright(B066+B067,2)) == 0 & ustrlen(B067) <= 5 & missing(B066,B067) == 0

replace 临时 = B066 + B067 if ustrpos(B066,"@") == 0 & ustrleft(B067,1) == "@" & B066 != ""
replace B066 = 临时 if 临时 != ""
replace B067 = "" if 临时 != ""

drop 临时
*****直接利用 id 内部现有数据进行修复
foreach i in B066 B067 {

```

```

by id (YEAR): replace `i' = `i'[_n-1] if missing(`i'[_n-1],`i'[_n+1]) == 0 & `i'[_n-1] == `i'[_n+1] &
`i' != `i'[_n-1] & 面板 id == 3 //根据上一观测值和下一观测值修复
foreach gsort in -YEAR YEAR { //必须采用[_n-1]形式, 而不是[_n+1]形式
    gsort id `gsort'
    by id: replace `i' = `i'[_n-1] if missing(`i'[_n-1],`i') == 0 & `i' != `i'[_n-1] &
(ustrregexra(`i',"(\s*[:punct:]+\s*\s)",".") == `i'[_n-1] | `i' == usubinstr(`i'[_n-1],".",",",.)) & 面板 id ==
3 //根据上一观测值修复 (考虑字符错误)
    by id: replace `i' = `i'[_n-1] if missing(`i'[_n-1],`i') == 0 & `i' != `i'[_n-1] &
strmatch(`i'[_n-1],ustrregexra(`i',"\\.??","*")) & 面板 id == 3 //根据上一观测值修复 (考虑字符遗漏)
}
}
*****批量修复为空值
foreach i in B066 B067 {
    replace `i' = ustrregexrf(`i',"^(http)?[:punct:]*(www)?[:punct:]*无[:punct:]*$", "") //将"无"替换为
空
    replace `i' = ustrregexrf(`i',"^([:punct:]+)$", "") //将纯符号替换为空
    replace `i' = ustrregexrf(`i',"^\u4e00-\u9fa5]+$", "") if ustrpos(`i',"网") == 0 & `i' != "有" //将纯汉
字替换为空值
    replace `i' = ustrregexrf(`i',"^\u4e00-\u9fa5]{2,}([a-z0-9\-\#]*\u4e00-\u9fa5)*$", "") if
ustrpos(`i',"网") == 0 //将类似于“北京市海淀区双清路?#楼?#室”的中文地址替换为空
    replace `i' = "" if real(`i') != . //将纯数字替换为空
    replace `i' = "" if strlen(`i') == 1 & `i' != "有" //将单字符替换为空
}
*****特殊修复
replace B067 = ustrregexra(B067,"^([0-9a-z]{8}[0-9x]|06|nog)$", "") //针对莱州市的一些错误情况
replace B066 = ustrregexra(B066,"bq[\u4e00-\u9fa5]+$", "") //针对中山市的一些错误情况, 注意与下
行的顺序
replace B066 = ustrregexra(B066,"y[0-9a-z]{8}[0-9x]$", "") //针对中山市的一些错误情况, 注意与上
行的顺序
replace B067 = ustrregexra(B067,"^(02)?y.*中山.*$", "") //针对中山市的一些错误情况
replace B067 = ustrregexra(B067,"^(08)?xz(06|07)?$", "") //针对高密市的一些错误情况
replace B067 = ustrregexra(B067,"^[0-9]{15}[0-9x]$", "") //针对镇江的一些错误情况
replace B067 = "" if B067 == "gy" //针对阜阳市的一些错误情况
replace B067 = "" if B067 == "dhgx_cld" //针对武汉东湖高新的一些错误情况

*****以下为 B066、B067 内容的互换
replace B067 = usubinstr(B067,".", "@", 1) if ustrregexm(B067,"w+\.\w+([:punct:]\s)\w+(\.w+)*$") &

```

```
ustrpos(B067,"www")==0
gen 临时 B066=""
gen 临时 B067=""
foreach i in B066 B067 { //针对 IP 地址为四组数字的情况
    replace `i'="http://" + ustrregexs(1) + "." + ustrregexs(2) + "." + ustrregexs(3) + "." + ustrregexs(4)
    if
    ustrregexm(`i',"^[a-z]{0,7}[:,punct:]\s*(\d{1,3})[:,punct:]\s*(\d{1,3})[:,punct:]\s*(\d{1,3})[:,punct:]\s*(\d{1,3})(:\d+)?(\\V.*)?$")
}
gen 互 换 标 记 = 1 if (B066 == "" |
ustrregexm(B066,"(^[\^@]*((http|www|http[:,punct:]\s*www)[\^@]*$)|(^http)[\^@]*(\d{1,3}\.){3}\d{1,3}(:\d+)?(\\V.*)?$)")) & (B067 == "" | ustrpos(B067,"@"))
replace 临时 B066 = B066 if 互换标记 == 1
replace 临时 B067 = B067 if 互换标记 == 1
replace B066 = 临时 B067 if 互换标记 == 1
replace B067 = 临时 B066 if 互换标记 == 1
drop 临时* 互换标记
save 互联网 1.dta, replace
```

*****二、批量处理（关于各种符号、http、www 等）

```
use 互联网 1.dta, clear
foreach i in B066 B067 {
    cap drop 修复前`i'
    gen 修复前`i'=`i',after(`i')
}
foreach i in B066 B067 {
    replace `i' = usubinstr(`i'," ",",",.)
    replace `i' = usubinstr(`i',"。",",",.)
    replace `i' = usubinstr(`i',"·",",",.) if strmatch(`i',"*@*")
    replace `i' = usubinstr(`i',"·",",",.) if strmatch(`i',"*@*") == 0 &
ustrlen(`i')-ustrlen(usubinstr(`i',"·",",",.)) >= 2 & strmatch(`i',"*w*w*w*") & ustrpos(`i',"") == 0 //注，
这里"·"出现两次以上，且可能包含"www"，形似 www.???com 类型网址
    replace `i' = usubinstr(`i',"",",",.) if strmatch(`i',"*@*,*@*") == 0 &
strmatch(`i',"*w*w*w*,*w*w*w*") == 0
    replace `i' = usubinstr(`i',"、",",",.) if strmatch(`i',"*@*、*@*") == 0 & strmatch(`i',"*w*w*w*、*w*w*w*") == 0
    replace `i' = ustrregexra(`i',"s\.\s*",".") //必须放在以上几条命令的后面
    replace `i' = ustrregexrf(`i',"?(=.\?\\V|.?.www|. *@)",".") if strmatch(`i',"*@*,*@*") == 0 &
```

5

```

replace `i' = ustrregexrf(`i',"^[htp]{2}.{0,5}www","http://www") if ustrpos(`i',"@") == 0
replace `i' = ustrregexs(2) if ustrregexm(`i',"^([[:punct:]]\s*)(http.*|www.*)$")

}

save 互联网 2.dta, replace

*****三、域名后缀的修复
use 互联网 2.dta, clear
foreach i in B066 B067 {
    cap drop 修复前`i'
    gen 修复前`i' = `i',after(`i')
}
*****顶级域名顺序打乱的修复
local 顶级域名 com net edu gov org biz //此处必须为 3 位的顶级域名, 否则下面的循环运算将会出错
foreach var in B066 B067 {
    foreach i of local 顶级域名 {
        local i 顺序重排
        foreach first of numlist 1/3 {
            local first = usubstr("`i'",`first',1)
            local i_first = usubinstr("`i'",`first'", "",1)
            foreach second of numlist 1/2 {
                local second = usubstr("`i_first'",`second',1)
                local i_second = usubinstr("`i_first'",`second'", "",1)
                foreach third of numlist 1 {
                    local third = usubstr("`i_second'",`third',1)
                    local 新增 i 顺序重排 = "`first`second`third'"
                    local i 顺序重排 : list i 顺序重排 | 新增 i 顺序重排
                }
            }
        }
        local regex = "(?<=\.)(\" + usubinstr("`i 顺序重排\"", " ", "|", .) + ")$"
        di "`regex'"
        replace `var' = ustrregexrf(`var',`regex',"`i'")
    }
}

```



```

        if "`j1'" != "`j2'" {
            local j1j2 = usubstr("`j1'",1,1) + usubstr("`j2'",2,1)
            di "`j1j2'"
            replace 混淆`var' = 1 if ustrright(`var',3) == "." + "`j1j2'"
        }
    }
}

****生成模糊顶级域名、模糊地区域名
foreach i of local 顶级域名 {
    local i 遗漏模糊组 `i'
    local i 模糊组 `i'
    forvalues order_i = 1/`=ustrlen("`i'")' {
        local i 遗漏模糊组 = "`i 遗漏模糊组'" + "|" + ustrleft("`i'",`order_i'-1) + "." +
usubstr("`i'",`order_i'+1,.)
        local i 模糊组 = "`i 模糊组'" + "|" + ustrleft("`i'",`order_i'-1) + "." +
usubstr("`i'",`order_i'+1,.)
    }
    di "`var'-----`i 遗漏模糊组'"
    replace `var' = ustrregexrf(`var',"(<!\w\\w.{0,100}|\..{0,3})\.(`i 遗漏模糊组')(?=\:|\V|$)",".`i'")
    foreach j of local 地区域名 {
        local j 模糊组 `j'
        forvalues order_j = 1/`=ustrlen("`j'")' {
            local j 模糊组 = "`j 模糊组'" + "|" + ustrleft("`j'",`order_j'-1) + "." +
usubstr("`j'",`order_j'+1,.)
        }
        di "`var'-----`i 模糊组'-----`j 模糊组'"
        replace `var' = ustrregexrf(`var',"\.(`i 模糊组')\.(`j 模糊组')[^a-zA-Z0-9\:\V]?(?=\:|\V|$)",".`i'.`j'") if strmatch(`var',"*www.?????") == 0 & 混淆`var' == 0
    }
}

}

drop 混淆*

*****cn 地区二级域名修复
local cn 地区二级域名 bj tj he sx nm ln jl hl sh js zj ah fj jx sd ha hb hn gd gx hi cq sc gz yn xz sn gs
qh nx xj hk mo tw //比如, bj.cn 为中国北京
foreach var in B066 B067 {
    replace `var' = ustrregexs(1) + "." + ustrregexs(2) + "." + ustrregexs(3) if

```

```

ustrregexm(`var',"^([^\-\\w@])('=substr("`cn 地区二级域名"," ","|",.)')([^\-\\w](cn(?:\d+)?(\V.*)?)$"))
}
//
replace B066 = ustrregexs(1) + "." + ustrregexs(2) + ".cn" if
ustrregexm(B066,"^([^\-\\w@])('=substr("`cn 地区二级域名"," ","|",.)')([^\-\\w](cn(?:\d+)?(\V.*)?)$"))
"","|",.)')([^\-\\w](cn(?:\d+)?(\V.*)?)$"))

*****特例修复
foreach var in B066 B067 {
    replace `var' = ustrregexs(1) + "." + ustrregexs(2) if
    ustrregexm(`var',"^([^\-\\w@])('=substr("`cn 地区二级域名"," ","|",.)')([^\-\\w](cn(?:\d+)?(\V.*)?)$"))
    replace `var' = ustrregexrf(`var',"\\..{0,2}0m(?:=|V|$)",".com")
    replace `var' = ustrregexs(1) if
    ustrregexm(`var',"^([^\-\\w@])('=substr("`cn 地区二级域名"," ","|",.)')([^\-\\w](cn(?:\d+)?(\V.*)?)$"))
    中国))[\u4e00-\u9fa5]+$")
}
replace B066 = substr(B066,".次年",".cn",.)
replace B067 =
ustrregexrf(B067,"[[:punct:]]s)+cn[^\-\\w]+(?:[ali]{1,3}.{1,2}[ba]{2,4}.?)[[:punct:]]s]*co[mn].?$",".cn.a
libaba.com")
replace B067 =
ustrregexrf(B067,"[[:punct:]]s]*cn[^\-\\w]*a?libab?a?[[:punct:]]s]*co[mn].?$",".cn.alibaba.com")
save 互联网 3.dta, replace

```

*****四、修复@

```

use 互联网 3.dta, clear
replace B067 = "" if B067 == B066 & ustrpos(B066,"@")
foreach i in B066 B067 {
    cap drop 修复前`i'
    gen 修复前`i' = `i',after(`i')
}
replace B066 = ustrregexs(1) if ustrregexm(B066,".*[^\-\\w@]{1,3}.{1,2}[ba]{2,4}.?") //提取各种括号里的
邮箱
*****利用变量组合修复
local i B066
foreach j in B067 {
    if "`i'" != "`j'" {
        foreach error in a 2 · # α . " " {

```

```

        di "`i';`j';`error'"
        replace                                `i'                                =
usubinstr(`i',"`error'+usubstr(`j',ustrpos(`j','.')+1,.),"@"+usubstr(`j',ustrpos(`j','.')+1,.)1) if `j' != "" &
ustrpos(`i',"@") == 0
    }
}
}
*****利用 id 组内修复
local i B066 //耗时较长
foreach gsort in -YEAR YEAR { //必须采用[_n-1]形式, 而不是[_n+1]形式
    gsort id `gsort'
    foreach error in a 2 · # α . " " {
        di "`i';`gsort';`error'"
        by id: replace `i' = `i'[_n-1] if missing(`i'[_n-1],`i') == 0 & `i' ==
usubinstr(`i'[_n-1],"@","`error'",.) & ustrpos(`i',"@") == 0 & 面板 id == 3
    }
}

*****利用知名邮箱修复
local i B066
replace `i' = usubinstr(`i',"@126.com","@126.com",.)
replace `i' = usubinstr(`i',"suhu.com","sohu.com",.)
replace `i' = ustrregexrf(`i',"@[^h]otmail|@h[^o]tmail|@ho[^t]mail","@hotmail")

local mail vip.163 vip.126 vip.sina vip.tom vip.qq vip.sohu alibaba-inc alibaba online 21cn yeah sohu
tom qq foxmail sina eyou hotmail outlook msn gmail yahoo netease 163 126 188 263 //顺序很重要, 辨
识度较强的放在前面
//修复知名邮箱多敲一个字符的情况
local 低辨识度 mail qq 163 126 263 188 //尤其是@163.com 邮箱, 其辨识度极低, 与@sh163.net、
@8163.net.cn 等邮箱容易混淆
local 高辨识度 mail : list mail - 低辨识度 mail
local regex = usubinstr(ustrregexra("`高辨识度 mail",".??",").*?(","(",")","|",",",.)
local regex = ustrregexra("`regex","\\((?=[^0-9a-z])",("\\") //注意需要两个\\, 因为第一个\\表示转义,
这里尤为特殊
local regex = usubstr("`regex'",2,ustrlen("`regex'")-2)
di "`regex'"
local expression ustrregexs(1)
forvalues order = 3/'=ustrlen("usbuinstr('高辨识度 mail'," ",",",",.)")+2' { //注意 ustrregexs(2)不计算

```

11

```

replace B066 = ustrregexrf(B066,"([:punct:]\s)*\@aliba?b?a?.*com$", "@alibaba.com") //alibaba.com
replace                                     B066                                     =
ustrregexrf(B066,"([:punct:]\s)*\@aliba?b?a?.*com([:punct:]\s)*cn$", "@alibaba.com.cn")
//alibaba.com
replace                                     B066                                     =
ustrregexrf(B066,"\@yah(\wo?o|o?o\w|\w?00|e{2}|u{1,2}|o{1,3}|w)([:punct:]\s)?(?=c)", "@yahoo.") if
ustrpos(B066,"@yahorng") == 0 & ustrpos(B066,"@yahg") == 0 & ustrpos(B066,"@yahua") == 0
//yahoo

```

```

save 互联网 4.dta, replace

```

*****五、其他修复

```

use 互联网 4.dta, clear

```

```

replace B067 = "" if B067 == B066 & ustrpos(B066,"@")

```

```

foreach i in B066 B067 {

```

```

    cap drop 修复前`i'

```

```

    gen 修复前`i' = `i', after(`i')

```

```

}

```

*****修复敲错一个字符的情况

```

foreach i in B066 B067 {

```

```

    gen 临时 = "" ,after(`i')

```

```

    bysort id (YEAR): egen 众数 = mode(`i')

```

```

    gen len = strlen(`i')

```

```

    sum len

```

```

    forvalues order_i = 1/'=r(max)' {

```

```

        replace 临时 = 临时 + " `order_i'" if usubstr(`i',`order_i',1) != usubstr(众数,`order_i',1) &
missing(B066,众数) == 0 & strlen(`i') == strlen(众数)

```

```

    }

```

```

    replace `i' = 众数 if wordcount(临时) == 1

```

```

    drop 临时 len 众数

```

```

}

```

*****利用邮箱域名修复网址

```

replace B067 = usubstr(B066,ustrpos(B066,"@")+1,.) if
ustrpos(usubstr(B066,ustrpos(B066,"@")+1,.),B067) & ustrregexm(B067,"^[a-z0-9\-\-]+$")

```

```

gen regex = "^(http.*|www.*)?" + ustrleft(B066,ustrpos(B066,"@")-1) +
"[:]punct:]\s)*(com|net|edu|gov|org|biz|cn|hk|tw|mo|jp|kr|uk|de|fr|it|es|pt|nl|ru|us|ca|cc| 中 国 )$" if

```

```

ustrpos(B066,"@") >= 2
replace B067 = ustrregexs(1) + "." + ustrregexs(3) if ustrregextm(B067,regex) & regex != ""
drop regex
*****

local                                website                                =
"(https?[:punct:]\s)*|www)[[:punct:]\s]*([0-9a-z\u4e00-\u9fa5][0-9a-z\u4e00-\u9fa5\-\.\?)*[:punct:]\s]
*((com|net|edu|gov|biz)?[:punct:]\s)*(com|net|edu|gov|org|biz|cn|hk|tw|mo|jp|kr|uk|de|fr|it|es|pt|nl|ru|us|c
a|cc|中国))(\.\d+)?(\.\?)*"

local                                email                                =
"([a-z0-9][a-z0-9\-\.\?)*\@([0-9a-z\u4e00-\u9fa5][0-9a-z\u4e00-\u9fa5\-\.\?)*[:punct:]\s]*((com|net|e
du|gov|biz)?[:punct:]\s)*(com|net|edu|gov|org|biz|cn|hk|tw|mo|jp|kr|uk|de|fr|it|es|pt|nl|ru|us|ca|cc|中国))"
//借助 B067 指标修复 B066 指标
gen 临时 B067 = ustrregexs(2) + "." + ustrregexs(3) if ustrregextm(B067,"^`website'$")
gen 临时 B066 = substr(B066,ustrpos(B066,"@")+1,..)
replace B066 = ustrleft(B066,ustrpos(B066,"@")) + 临时 B067 if ustrpos(临时 B067,临时 B066) == 1
& ustrpos(B066,"@")
drop 临时*
//借助 B066 指标修复 B067 指标
gen 临时 B066 = ustrregexs(2) + "." + ustrregexs(3) if ustrregextm(B066,"^`email'$")
gen 临时 B067 = substr(B067,ustrpos(B067,".")+1,..)
replace B067 = ustrleft(B067,ustrpos(B067,".")) + 临时 B066 if ustrpos(临时 B066,临时 B067) == 1 &
ustrpos(B067,".")
drop 临时*
//单个指标两个值的情况
replace B067 = ustrregexs(1) + "或者" + ustrregexs(9) if ustrregextm(B067,"^(`website')[:punct:]\s\或\
和]*(`website')$")
replace B066 = ustrregexs(1) + "或者" + ustrregexs(7) if ustrregextm(B066,"^(`email')[:punct:]\s\或\
和]*(`email')$")

*****个别修复
replace B066 = ustrregexrf(B066,"^转?\d+[-?\d+$","") //手机或电话号码替换为空
replace B067 = ustrregexrf(B067,"^[u4e00-\u9fa5]{2,}\s*[a-z]\d{2,}$","") //将“美国 h502”这种类型
替换为空值
replace B067 = ustrregexrf(B067,"^.*co.ltd$","") //
replace B067 = ustrregexrf(B067,"^y[u4e00-\u9fa5]*$","")
replace                                B067                                =
ustrregexrf(B067,"([:punct:]\s)*cn\.(ali?b?a?b?a?[alibaba]{6,8})\.\.?c?o?m?$",".cn.alibaba.com") // 针
对 cn.alibaba.com 的修复

```

```

foreach var in B066 B067 {
    replace `var' = "" if `var' == ustrleft(`var',1)*ustrlen(`var') & ustrlen(`var') >= 2 & `var' != "www"
    replace `var' = "" if ustrregexm(`var',"^\w+[\u4e00-\u9fa5]+$") & ustrpos(`var',"网") == 0
    replace `var' = "" if ustrregexm(`var'," (电话|传真)")
}

*****直接利用 id 内部现有数据进行修复
replace B067 = "http://" + B067 if strmatch(B067,"www.*")
foreach i in B066 B067 {
    by id (YEAR): replace `i' = `i'[_n-1] if missing(`i'[_n-1],`i'[_n+1]) == 0 & `i'[_n-1] == `i'[_n+1] &
    `i' != `i'[_n-1] & 面板 id == 3 //根据上一观测值和下一观测值修复
    foreach gsort in -YEAR YEAR { //必须采用[_n-1]形式, 而不是[_n+1]形式
        gsort id `gsort'
        by id: replace `i' = `i'[_n-1] if missing(`i'[_n-1],`i') == 0 & `i' != `i'[_n-1] &
        (ustrregexra(`i',"(\s*[:punct:]+\s*\s)",".") == `i'[_n-1] | `i' == usubinstr(`i'[_n-1],".",",",.)) & 面板 id
        == 3 //根据上一观测值修复 (考虑字符错误)
        by id: replace `i' = `i'[_n-1] if missing(`i'[_n-1],`i') == 0 & `i' != `i'[_n-1] &
        strmatch(`i'[_n-1],ustrregexra(`i',"?.?",*)) & 面板 id == 3 //根据上一观测值修复 (考虑字符遗漏)
    }
}

*****以下为 B066、B067 内容的互换
replace B067 = usubinstr(B067,".", "@",1) if ustrregexm(B067,"^\w+\.\w+([:punct:]\s)\w+(\.w+)?$") &
ustrpos(B067,"www") == 0
gen 临时 B066 = ""
gen 临时 B067 = ""
replace 临时 B066 = B066 if ustrpos(B066,"@") == 0 & ustrpos(B067,"@")
replace 临时 B067 = B067 if ustrpos(B066,"@") == 0 & ustrpos(B067,"@")
replace 临时 B066 = B066 if ustrpos(B066,"@") == 0 & strmatch(B066,"*http*www*") & B067 == ""
replace 临时 B067 = B067 if ustrpos(B066,"@") == 0 & strmatch(B066,"*http*www*") & B067 == ""
replace B066 = 临时 B067 if 临时 B066 != "" | 临时 B067 != ""
replace B067 = 临时 B066 if 临时 B066 != "" | 临时 B067 != ""
drop 临时*

save 互联网 5.dta, replace

*****

use 互联网 5.dta, clear

```

keep if YEAR >= 2001 & YEAR <= 2009 //注: 本应该是 2004-2009 年, 但是经过多轮修复后, 已将网址数据扩展到 2001 年

```
replace B067 = "" if ustrpos(B067,"add")
```

```
replace B067 = "" if strpos(B066,B067) == 0 & strlen(B067) <= 3
```

```
replace B067 = "" if ustrregexm(B067,"^[\\d\\s]+$")
```

```
//
```

```
gen 网址哑变量 = .
```

```
replace 网 址 哑 变 量 = 1 if  
ustrregexm(B067,"\\.(com|net|edu|gov|org|biz|cn|hk|tw|mo|jp|kr|uk|de|fr|it|es|pt|nl|ru|us|ca|cc|  
国)(\\.\\d+)?(\\.\\*)?$") 中
```

```
replace 网址哑变量 = 1 if ustrregexm(B067,"^[:punct]]*(https?|ftp|www)")
```

```
replace 网址哑变量 = 1 if ustrregexm(B067,"\\d+\\.\\d+\\.\\d+\\.\\d+")
```

```
replace 网址哑变量 = 1 if ustrpos(B067,"网") //针对“中国金属门供应网”这种网址
```

```
replace 网址哑变量 = 1 if B067 == "有"
```

```
replace 网址哑变量 = 0 if 网址哑变量 != 1
```

```
//
```

```
gen 邮箱哑变量 = .
```

```
replace 邮箱哑变量 = 1 if ustrpos(B066,"@")
```

```
replace 邮箱哑变量 = 0 if 邮箱哑变量 != 1
```

```
//
```

```
gen internet = (邮箱哑变量 == 1 | 网址哑变量 == 1)
```

```
drop 修复前*
```

```
save 互联网 f.dta, replace
```

```
sum YEAR if inrange(YEAR,2001,2009)
```

```
local num = r(N)
```

```
sum YEAR if inrange(YEAR,2001,2009) & 备份 B066 != B066
```

```
di r(N)/`num'
```

```
gen old 邮箱哑变量 = ustrregexm(备份 B066,"@")
```

```
sum YEAR if 邮箱哑变量 != old 邮箱哑变量
```

```
di r(N)/`num'
```

```
sum YEAR if inrange(YEAR,2004,2009)
```

```
local num = r(N)
```

```
sum YEAR if inrange(YEAR,2004,2009) & 备份 B067 != B067
```

```
di r(N)/`num'
```

```
gen old 网址哑变量 = ustrregexm(备份 B067,"www|WWW|com|COM|cn|CN|http|HTTP")
```

```
sum YEAR if 网址哑变量 != old 网址哑变量
```

$$\frac{dr(N)}{N}$$

附录 2: stata 程序之园区哑变量生成

```
local 变量组 B051 B053 B054 B052 B50CHS // 乡镇 街道 [村居]委会 村(街、门牌号) 详细地址
local 辅助变量组 B056 B057 B065 //省地县码 乡村码 邮编
clear all
save 地址.dta, replace emptyok
local a : dir . files "a_*.dta"
foreach i of local a {
    use `i', clear
    foreach var in `变量组' `辅助变量组' {
        cap gen `var' = ""
    }
    keep id YEAR FRDM B02 `变量组' `辅助变量组'
    append using 地址.dta
    save 地址.dta, replace
}
order id YEAR FRDM B02 `变量组' `辅助变量组'
//全角转半角等修复
foreach var of local 变量组 {
    gen 备份`var' = `var'
    replace `var' = ustrlower(ustrtrim(`var'))
    foreach j of numlist 33/64 91/126 {
        replace `var' = usubinstr(`var',uchar(`j'+65248),uchar(`j'),.) //全角和半角相差 65248
    }
    replace `var' = usubinstr(`var',"—","-",.) //这两个不是全角和半角的区别
    replace `var' = ustrregexra(`var',"\\s|null","")
    replace `var' = ustrregexra(`var',"o","O")
    replace `var' = "" if ustrregexm(`var',"^[0-9]+$|^0\\.00$|^[a-z]+$|e\\+")) //e+是科学计数
}

merge m:1 id using 面板数据 id 提取.dta,keep(1 3) generate(面板 id)
save 地址.dta, replace

*****大循环修复(主要针对 B051、B052)
//ssc install unique
use 地址.dta, clear

local 变量组 B051 B053 B054 B052
```

```
local 关键词 ((?<!(^集)镇|乡|民族乡|苏木|民族苏木)(\[([\u4e00-\u9fa5]+[\ ])?)?$ (街道?办事?处?|办
事处?|地区办事处|虚拟街道?|街道|地区|(?<!(居民)管理?委员?会?(筹备)?工作?委员?会?|园区|(经开|
火炬|保税|贸易|度假|旅游|风景)区|科技园|软件[谷园])(示范|高新|工业|开发|加工|产业|物流)[园
区][农林牧渔]场)(\[([\u4e00-\u9fa5]+[\ ])?)?$ ((社区|虚拟|虚设|类似|村)?([居村]民?委?员?会?)(社
区)?居民管理委|(社区)?(家属?委员?会?|家属))(嘎查)?(牧民?委员?会?)(?<=[〇一二三四五六七八
九十百零壹贰叁肆伍陆柒捌玖拾佰两 0-9])委|社区|行政村|嘎查|生产队|连队|队|生活
区)(\[([\u4e00-\u9fa5]+[\ ])?)?$ ((?<!(行政|虚拟)村庄|寨|沟|洼|屯|营|(?<=[〇一二三四五六七八九十
百零壹贰叁肆伍陆柒捌玖拾佰两 0-9a-z])(号|区|座|幢|栋|楼|单元|层|室|房|
米)))(\[([\u4e00-\u9fa5]+[\ ])?)?$
```

```
foreach var of local 变量组 {
```

```
    qui gen 修复前`var' = `var', after(`var')
```

```
    replace `var' = ustrregexra(`var',"^[^w[:punct:]]","") //乱码先替换为空值, 不必担心, 因为下面
    的循环将会对已替换为空值的乱码进行填充式修复, 修复为非乱码字符
```

```
}
```

```
foreach 大循环 in 1 2 {
```

```
    di "-----****大循环第`大循环'次****-----"
```

```
    **** (一) B051、B053 的修复, 主要是字符遗漏。比如。良乡镇写成了良乡, 黄村镇写成了
    了黄村, 西乡街道写成了西乡。
```

```
    foreach var in B051 B053 B054 B052 {
```

```
        local var 关键词 : word `list posof "`var'" in 变量组' of `关键词'
```

```
        local change .
```

```
        while `change' > 0 { //可以设定终止条件, 0 是最严的, 也可以设置一个比较小的正数
```

```
            di "-----`var'-----"
```

```
            qui gen compare = `var'
```

```
            **利用 B056、YEAR 分组修复 B051 和 B053
```

```
            if "`var'" == "B051" | "`var'" == "B053" {
```

```
                qui gen `var'纯名称 = ustrregexra(`var',"`var' 关键词',"")
```

```
                qui unique `var', by(YEAR B056 `var'纯名称) generate(`var'名称非重复个数) //空
                值也会被按照非重复值统计进去
```

```
                qui bysort YEAR B056 `var'纯名称 (`var'名称非重复个数) : replace `var'名称非重
                复个数 = `var'名称非重复个数[_n-1] if `var'名称非重复个数 == .
```

```
                gsort YEAR B056 `var'纯名称 -`var'
```

```
                by YEAR B056 `var'纯名称: replace `var' = `var'[_n-1] if missing(YEAR,B056,`var'
                纯名称) == 0 & strmatch(`var'[_n-1],ustrregexra(`var',"?.?","*")) & `var'名称非重复个数 == 2 //根据
                上一观测值修复 (考虑字符遗漏)
```

```
                drop `var'纯名称 `var'名称非重复个数
```

```

    }
    **利用 id、YEAR 分组修复
    bysort id (YEAR): replace `var' = `var'[_n-1] if missing(`var'[_n-1],`var'[_n+1]) == 0 &
`var'[_n-1] == `var'[_n+1] & `var' != `var'[_n-1] & ustrregexm(`var'[_n-1],"var 关键词") & 面板 id ==
3 //根据上一观测值和下一观测值修复。注: 添加"面板 id == 3"这个条件能够保证 while 循环不陷
入死循环。下同。
    foreach gsort in YEAR -YEAR { //必须采用[_n-1]形式, 而不是[_n+1]形式
        gsort id `gsort'
        by id : replace `var' = `var'[_n-1] if missing(`var'[_n-1],`var') == 0 & `var' !=
`var'[_n-1] & strmatch(`var'[_n-1],ustrregexra(`var',"??","*")) & ustrregexm(`var'[_n-1],"var 关键词")
& 面板 id == 3 //根据上一观测值修复(考虑字符遗漏)
    }
    replace `var' = ustrregexrf(`var',"^([\u4e00-\u9fa5]{2,3}(维吾尔族?|回族|壮族)?(省|自治
区)?([\u4e00-\u9fa5]{2,4}(市|盟)|[\u4e00-\u9fa5]{2,8}(自治州))?([\u4e00-\u9fa5]{2,4}((?!经开|火
炬|保税|度假|旅游|风景|示范|高新|工业|开发|加工|产业|管理|社|地)区|县|市|([\u4e00-\u9fa5]族)*自
治县|旗))?(?!镇|乡|办事处|街道|社区|地区|虚拟|虚设|类似|居民?委|居民管理委|村民?委|牧民?委|家
属?委|工作?委|管理?委|委|)|)",") //由于中国的地级行政单位的"地区"与乡级行政单位的"地区"几
乎无法区分开, 所以地级行政单位的正则法则采用了"(市|盟|自治州)", 而非"(市|盟|自治州|地区)"
    sum YEAR if `var' != compare
    local change = r(N)
    drop compare
}
}
save 地址 1.dta, replace
**** (二) 关于内容重复&内容臃肿的修复
forvalues order_i = 1/`=wordcount("`变量组'") {
    forvalues order_j = 1/`=wordcount("`变量组'") {
        local i : word `order_i' of `变量组'
        local i 关键词 : word `order_i' of `关键词'
        local j : word `order_j' of `变量组'
        local j 关键词 : word `order_j' of `关键词'
        qui gen i 匹配组别 = 1 if ustrregexm(`i',"i 关键词") //由于下面的循环用到, 所以需要
要提前生成
        qui gen j 匹配组别 = 1 if ustrregexm(`j',"j 关键词") //由于下面的循环用到, 所以需要
要提前生成
        if `order_i' != `order_j' {
            di "利用 `i'、`j'进行重复修复"
            replace `i' = "" if strmatch(`j',ustrregexra(`i',"??","*")) & missing(`i',j,j 匹配组别)

```

== 0 & missing(i 匹配组别) //内容重复修复。

```

    }
    if `order_i' < `order_j' {
        di "利用`i'、`j'进行臃肿修复"
        replace `j' = substr(`j',ustrpos(`j',`i')+ustrlen(`i'),.) if ustrpos(`j',`i') & missing(`i',`j',i
匹配组别) == 0 & strmatch(`j',"*(`i')*") == 0 &
ustrregexm(substr(`j',ustrpos(`j',`i')+ustrlen(`i'),.),"^j 关键词") == 0 //注: 最后一个条件是为了防止
出现类似于"圪洞镇居委会"这种特殊情况。倒数第二个条件是为了防止回避类似于"高立庄村(花
乡构件厂墙外)"。
    }
    drop *匹配组别
}
}

```

save 地址 1.dta, replace

**** (三) 关于内容张冠李戴的修复(在理论上, 下面的内容张冠李戴循环与上面的内容重复 & 内容臃肿循环, 最好不要合并在同一个循环运算中, 这是因为上面的循环与下面的循环都涉及多组两两变量)

```

forvalues order_i = 1/'=wordcount("`变量组'")' {
    forvalues order_j = `order_i'+1/'=wordcount("`变量组'")' {
        local i : word `order_i' of `变量组'
        local i 关键词 : word `order_i' of `关键词'
        local j : word `order_j' of `变量组'
        local j 关键词 : word `order_j' of `关键词'
        qui gen `i'匹配组别 = 1 if ustrregexm(`i',"i 关键词") //因为每一次循环都可能会使匹
        配组别发生变化, 所以每次循环前都要重置
        qui gen `j'匹配组别 = 1 if ustrregexm(`j',"j 关键词") //因为每一次循环都可能会使匹
        配组别发生变化, 所以每次循环前都要重置
        qui gen `i'和`j'互换标记 = 0
        qui replace `i'和`j'互换标记 = `i'和`j'互换标记 + 1 if (ustrregexm(`i',"j 关键词") &
        missing(`i'匹配组别)) | `i' == ""
        qui replace `i'和`j'互换标记 = `i'和`j'互换标记 + 1 if (ustrregexm(`j',"i 关键词") &
        missing(`j'匹配组别)) | `j' == ""
        qui replace `i'和`j'互换标记 = 0 if `i' == "" & `j' == ""
        qui gen 临时`i' = `i' if `i'和`j'互换标记 == 2
        qui gen 临时`j' = `j' if `i'和`j'互换标记 == 2
        replace `i' = 临时`j' if `i'和`j'互换标记 == 2
        replace `j' = 临时`i' if `i'和`j'互换标记 == 2
        drop 临时* *匹配组别 *互换标记
    }
}

```

```

    }
  }
  save 地址 1.dta, replace
  **** (四) 标准化处理

  foreach var in B051 B053 {
    replace `var' = ustrregexrf(`var',"(?<=[\u4e00-\u9fa5]{1,100}[镇乡])(政府办公室|政府|人民政府|政府办事处)(?=$\()","")

    replace `var' = ustrregexrf(`var',"(?<=[\u4e00-\u9fa5]{2,100})(街道?办事?处?|办街道办事处|办街道事处|街道办办事处|街道)(?=$\()","街道办事处")

    replace `var' = ustrregexrf(`var',"(?<=[\u4e00-\u9fa5]{2,100})地区办事(?=$\()","地区办事处")
  }
  //

  foreach var in B054 {
    replace `var' = ustrregexrf(`var',"居民?委?员?会?)(?=$\()","居委会")
    replace `var' = ustrregexrf(`var',"村民?委?员?会?)(?=$\()","村委会")
    replace `var' = ustrregexrf(`var',"家属?委员?会?|家属)(?=$\()","家委会")
    replace `var' = ustrregexrf(`var',"牧民?委员?会?)(?=$\()","牧委会")
    replace `var' = ustrregexs(1) + "第" + ustrregexs(2) + "居委会" + ustrregexs(3) if
    ustrregexm(`var',"(^.*)第?([〇一二三四五六七八九十百零壹贰叁肆伍陆柒捌玖拾佰两〇-9]+)委
    ($\()") & ustrpos(B053,"街道")

    foreach 辅助 var in B051 B053 {
      bysort YEAR B056 `辅助 var' (`var') : replace `var' = `var'[_n-1] if
      ustrregexrf(`var',"(?<=[2,20})(乡|镇|街道(办事处)?|地区办事处)(?=(居委会|村委会|家委会|牧委会)($\()","") == `var'[_n-1] & ustrpos(`var',`辅助 var') == 1

      bysort YEAR B056 (`var') : replace `var' = `var'[_n-1] if
      ustrregexrf(`var',"(?<=[2,20})(乡|镇|街道(办事处)?|地区办事处)(?=(居委会|村委会|家委会|牧委会)($\()","") == `var'[_n-1] & ustrpos(`var',`辅助 var') == 1

      bysort YEAR id (`var') : replace `var' = `var'[_n-1] if ustrregexrf(`var',"(?<=[2,20})(乡|
      镇|街道(办事处)?|地区办事处)(?=(居委会|村委会|家委会|牧委会)($\()","") == `var'[_n-1] &
      ustrpos(`var',`辅助 var') == 1
    }
  }
  //

  foreach var in B051 B053 B054 B052 {
    replace `var' = ustrregexrf(`var',"(?<=[\u4e00-\u9fa50-9]{2,100})(管理?委员?会?)","管委会")

    replace `var' = ustrregexrf(`var',"(?<=[\u4e00-\u9fa50-9]{2,100})(工作?委员?会?)","工委会")
  }

```

```

")
    }
    save 地址 1.dta, replace
}
drop 修复前*
compress
save 地址 1.dta, replace

****B054、B052 的修复
use 地址 1.dta, clear
gen 修复前 B054 = B054, after(B054)
foreach i in B051 B053 {
    bysort YEAR B056 `i' (B054) : replace B054 = B054[_n-1] if
ustrregexrf(B054,"(?<=. {2,20})(乡|镇|街道(办事处)?|地区办事处)(?=(居民委员会|村民委员会|家属
委员会|牧民委员会|居委会|村委会|家委会|牧委会)($\|())", "") == B054[_n-1] & ustrpos(B054, `i') ==
1

    bysort YEAR B056 (B054) : replace B054 = B054[_n-1] if ustrregexrf(B054,"(?<=. {2,20})(乡|镇|
街道(办事处)?|地区办事处)(?=(居民委员会|村民委员会|家属委员会|牧民委员会|居委会|村委会|家
委会|牧委会)($\|())", "") == B054[_n-1] & ustrpos(B054, `i') == 1

    bysort YEAR id (B054) : replace B054 = B054[_n-1] if ustrregexrf(B054,"(?<=. {2,20})(乡|镇|
街道(办事处)?|地区办事处)(?=(居民委员会|村民委员会|家属委员会|牧民委员会|居委会|村委会|家
委会|牧委会)($\|())", "") == B054[_n-1] & ustrpos(B054, `i') == 1
}
//修复 B052 中的中文数字
foreach var in B054 B052 {
    cap gen 修复前 `var' = `var', after(`var')
    qui gen 不修改标记 `var' = ustrregexm(`var', "[〇一二三四五六七八九零壹贰叁肆伍陆柒捌玖
两][0-9][0-9][〇一二三四五六七八九零壹贰叁肆伍陆柒捌玖两][十拾百佰千仟]{2,}")
    replace `var' = ustrregexra(`var', "[\s]", "") //空格替换为空非常重要, 因为后面将会构造空格作分
割。
    replace `var' = ustrregexra(`var', "(?<![〇一二三四五六七八九零壹贰叁肆伍陆柒捌玖两
0-9][0-9][十拾]", "一十") if 不修改标记 `var' != 1 //关于 10 的修复
    replace `var' = ustrregexra(`var', "(?<=[〇一二三四五六七八九零壹贰叁肆伍陆柒捌玖两 0-9 万
亿])[〇零]^[〇零]${[〇零]}(?=[〇一二三四五六七八九零壹贰叁肆伍陆柒捌玖两 0-9 万亿])", "0") if
不修改标记 `var' != 1 //关于 0 的修复
    local 非零中文数字 一壹 二贰两 三叁 四肆 五伍 六陆 七柒 八捌 九玖
    local 非零阿拉伯数字 1 2 3 4 5 6 7 8 9
    forvalues order = 1/`=wordcount("非零中文数字")' { //关于 1-9 的修复

```

```
replace `var' = ustrregexra(`var',"(<=[第道路街巷弄 a-z、\#-\—\+])[\〇零十拾百佰千仟
`=usubinstr("`非零中文数字"," ",",",.)'0-9]{0,18})['=word("`非零中文数字`,`order')][`=word("`非零
中文数字`,`order')](?=[\〇零十拾百佰千仟`=usubinstr("`非零中文数字"," ",",",.)'0-9]{0,18}[号桥道
路街巷弄段条区座幢栋楼单元层室间门户房米组排队社军师旅团连班 a-z、\#-\—\+])",`=word("`
非零阿拉伯数字`,`order')") //{0,18}中参数选 18 已经足够, 即使身份证号也没超过 18 位。
```

```

}
//
qui gen 数字算式 = ustrregexra(`var',"[^0-9 十拾百佰千仟]+"," ") if 不修改标记 `var' !=
1,after(`var')
qui replace 数字算式 = ustrregexra(数字算式,"(?<=^\s)([十拾百佰千仟]+[0-9]+)(?<=s$)"," ") //
对于类似于"千百"这种类型, 以及不含"千百十"的纯数字, 没必要浪费计算机内存进行修复
qui gen 数字算式 count = wordcount(数字算式) if 不修改标记 `var' != 1
sum 数字算式 count
forvalues order = 1/`=r(max)' {
    di "`order'"
    qui gen 数字算式`order' = word(数字算式,`order')
    qui replace 数字算式`order' = ustrleft(ustrregexs(1),1) + "千" + ustrleft(ustrregexs(2),1) + "
百" + ustrleft(ustrregexs(3),1) + "十" + ustrleft(ustrregexs(4),1) if ustrregexm(word(数字算
式,`order'),"([1-9][千仟])?([1-9][百佰])0?([1-9][十拾])0?([1-9])?") & strlen(数字算式`order') !=
strlen(数字算式`order') //正则条件中必须保证至少出现一次千仟百佰十拾的关键词, 所以利用汉字的
的 strlen 和 strlen 不同这个属性
    qui replace 数字算式`order' = ustrregexra(数字算式`order',"^(?<=千)(?<=千)(?<=百)(?<=
百)(?<=十)(?<=十)$","0")
    qui replace 数字算式`order' = stofreal(real(ustrregexra(数字算式`order',"[千百十]", "")))
    replace `var' =usubinstr(`var',word(数字算式,`order'),数字算式`order',1)
}
drop 数字算式* 不修改标记*
}
**
drop 修复前*
compress
save 地址 2.dta, replace

****
use 地址 1.dta, clear
keep if YEAR == 2010
save 临时.dta, replace

```

```
use 地址 1.dta, clear
gen park = 0 if (inrange(YEAR,1996,2009) | inrange(YEAR,2011,2013))
replace park = 1 if ustrregexm(B051+B053+B054+B052,"(园区|(经开|火炬|保税)区|科技园|软件[谷园])|(示范|高新|工业|开发|加工|产业)[园区]|科学|技术|科创|创新)") & (inrange(YEAR,1996,2002) | inrange(YEAR,2004,2009))
replace park = 1 if ustrregexm(备份 B051+备份 B053+备份 B054+备份 B052,"(园区|(经开|火炬|保税)区|科技园|软件[谷园])|(示范|高新|工业|开发|加工|产业)[园区]|科学|技术|科创|创新)") & (inrange(YEAR,1996,2002) | inrange(YEAR,2004,2009))
replace park = 1 if ustrregexm(B50CHS,"(园区|(经开|火炬|保税)区|科技园|软件[谷园])|(示范|高新|工业|开发|加工|产业)[园区]|科学|技术|科创|创新)") & (YEAR == 2003 | YEAR == 2008 | inrange(YEAR,2011,2013))
replace park = 1 if ustrregexm(B057,"^[1-6][0-9]{3}$") & inrange(YEAR,1996,2002)
drop if YEAR == 2010
bysort id (YEAR) : replace park = 1 if park[_n-1] == 1 & park[_n+1] == 1 & park == 0 & (inrange(YEAR,1996,2009) | inrange(YEAR,2011,2013))
append using 临时.dta
sort id YEAR
save 地址 f.dta, replace
```

附表 1

附表 1

web、mail 数据详细修复及示例

修复归类	说明	示例
小写化 web 和 mail。	URL 中的协议、主机名、端口等，以及绝大多数邮箱的用户名一般不区分大小写。	比如，将“Http”“HTTP”“http”“hTTp”修复为“http”。
半角化 web 和 mail。	URL 中的协议、主机名、端口等，以及绝大多数邮箱的用户名一般只支持半角。	比如，将“@126. c o m”修复为“@126.com”。
将无效的 web 和 mail 替换为空值。	1. 原数据以某些显性标记将 web、mail 标注为空值的情况。 2. 原数据中的 web、mail 被错填为非企业网址、非企业邮箱数据的情况。	1. 比如，把 web 和 mail 中标注为“无”“null”“0”的值替换为空值。 2. 比如，莱州宏利非织布有限公司 ^① 在 2006 年的 web 被错填为原组织结构代码“16987458X”；中山市汽车大修厂有限公司 ^② 在 2005 年的 mail 被错填为带前缀“y” ^③ 的原组织结构代码“y198077387”。 比如，广东海灵保健制品有限公司在 2006 年的 web 被错填为带前缀“y”的、残缺的原企业名称“y 中山市海灵保健制品有”；武汉精一印刷有限公司的 web 被错填为英文名称“wuhanreliance printing.co.ltd”。 比如，莱州市很多企业的 web 被错填为地级区划代码“06” ^④ ；上海市很多企业在 2007 年的 web 被错填为“台湾 h143”“香港 h110”“德国 h304”“美国 h502”“日本 h116”“意大利 h307”等特殊代码。 比如，武汉东湖高新区很多企业的 web 被错填为 dhgx_cld（前 4 位为“东湖高新”拼音首字母）；中山市的很多企业的 mail 被错填为“bq 石岐区”“bq 东升镇”“bq 板芙镇”“bq 小榄镇”“bq 东区”等 ^⑤ 。
利用正则法则及逻辑判断去批量修复或人工修复 web 和 mail 中的 typo 错误。	1. 可利用多套正则法则、逻辑判断组合去批量修复的情况。 2. 需要人工修复的情况。	1. 比如，把“@”“◎”“.”“#”“a”“α”“2” ^⑥ 等修复为“@”，把“@yahu”“@yahuo”“@yahuoo”“@yahooo”等错误的雅虎邮箱修复为“@yahoo”，把“@21cn.not”“@21cnnet”“@21cn.ne”“@21.cn.net”“@21cn.et”“@21.cnt”“@21.cn.nel”等错误的 21CN 邮箱修复为“@21cn.net”，把“hbbp://”“hhttp:\\\\”“hhttp//”“http://”“rttp:11”“http:111”等错误的协议修复为“http://”，把“.nte”“.ten”“.ent”等错误的顶级域名修复为“.net”，把“.cn.alibaba.c”“”，cn，

① 新组织机构代码为“757462282”。

② 新组织机构代码为“L09140061”。

③ 本文猜测“y”可能表示原来、原先等意思。

④ 判断依据：莱州县级市属于烟台地级市代管，而山东省烟台市的区划代码为“3706”。

⑤ 注意，由于中山市是不设市辖区的特殊地级市，所以“石岐区”和“东区”都属于乡级行政单位，而非县级行政单位。

⑥ 因为 Shift 键加数字键“2”为“@”。

		alibaba.” “.cn.al.baba.com” “.cn.alibuba.com” 等错误的阿里巴巴旺铺修复为 “.cn.alibaba.com”, 等等。 2. 比如, 把 “次年” 修复为 “cn”, 把 “bjcn” (北京)、“tjcn” (天津)、“hecn” (河北)、“sxcn” (山西)、“nmcn” (内蒙古) 等错误的中国省级域名修复为 “cn” 前带符号 “.” 的形式, 等等。
利用跨年数据或其它指标数据协助修复 web 和 mail 的漏报漏录。	1. 可利用跨年数据进行修复的情况。 2. 可利用其他指标数据协同修复的情况。	1. 比如, 内蒙古塞飞亚集团有限责任公司 2004—2006 年的 web 分别为 “http://www.saifeiya.com”、空值、“www.saifeiya.com”, 所以应该将 2005 年漏报漏录的 web 填充为 “http://www.saifeiya.com” (此处自动修复为带协议的 URL 形式)。 比如, 福建省武夷酒业有限公司 2007—2009 年的 mail 分别为 “wyjy199@vip.163.com” “null” “wyjy199@vip.163.com”, 所以可以将 2008 年的空值标记 “null” 替换为 “wyjy199@vip.163.com”。 2. 比如, 沈阳石蜡化工有限公司 2002 年的 web 为空值, 但该年的 mail 却被错误地录入为 “www.sypc.c”, 所以可以从 mail 中提取出关键域名 “sypc”, 然后利用正则法则、逻辑判断等组合, 将其修复为 “www.sypc.com.cn” 并填充原 web 的空值。 比如, 嘉兴锦林木业有限公司 2003 年的 web 为空值, 但该年的 mail 却被错误地录入为 “jinlumber@sina.comwww.jinlinlumber.com”, 所以可以从 mail 中分离出 “www.jinlinlumber.com” 并填充原 web 的空值。
利用中国工业企业数据库中多个指标协助修复 web 和 mail 的混报混录。	1. “张冠李戴型” 混报混录的情况。 2. “一拆多型” 混报混录的情况。 3. “多合一型” 混报混录的情况。 4. “复杂错误型” 混报混录的情况。	1. 比如, 天津市大港区靖元机械制造有限公司 2008 年的 web、mail 被混报混录为 “jy@tjyingyuan.com” “http://www.tjyingyuan”, 所以应该先将 web、mail 值调换, 再分别将其修复为 “http://www.tjyingyuan.com” (此处利用正则法则、逻辑判断等组合自动修复协议遗漏的 “:” 及后缀遗漏的 “.com”)、“jy@tjyingyuan.com”。 2. 比如, 达州市新建水泥有限公司 2008 年的 web、mail 被混报混录为 “com” “sinjianshuini2163”, 所以应该先将 mail、web 值合并 (此处综合利用正则法则、逻辑判断等确定 mail、web 的合并顺序), 再将 web 和 mail 分别修复为空值、“sinjianshuini@163.com” (此处综合利用正则法则、逻辑判断等自动将 “2” 修复为 “@”)。 3. 比如, 南昌汽缸垫厂 2003 年的 web、mail 被混报混录为空值、“www.cm-gasket.comncqgdswy@nc.jx.cn”, 所以需要先将 mail 值分拆为 “www.cm-gasket.com” “ncqgdswy@nc.jx.cn”, 再根据不同的正则法则分别将 web、mail 修复为 “www.cm-gasket.com” “ncqgdswy@nc.jx.cn”。 4. 比如, 开平嘉达摩托车配件制品有限公司 2007 年的 web 被错误地记录为三位乡级代码 “109”^①, 所以需要先将 web 替换为空, 再结合正则法则、逻辑判断以及其他多个指标, 将其修复为 “http://www.kpgada.com”。

附表 2

① 判断依据: 在补上开平县级市的六位县级代码之后, 可得完整的区划代码 “440783109”, 恰好为企业的所在地——开平市赤坎镇

附表 2

基于 web 生成 internet 变量及示例

归类	示例
web 中的主机名含有“com”“net”“edu”“gov”“org” “biz”“cn”“中国”“hk”“tw”“mo”“jp”“kr” “uk”“de”“fr”“it”“es”“pt”“nl”“ru”“us” “ca”“cc”等顶级域名。	比如, 昭和沥青工业(上海)有限公司的网址 “shoreki.co.jp”含有日本顶级域名“jp”, 所以 可判定其 internet 取值为 1。
web 中的主机名含有“www”关键词。	比如, 云南省玉溪印刷有限责任公司的网址被错 填为“w.w.w.yuxiprinting.com”, 本文将其修复 为“http://www.yuxiprinting.com”, 由于含有 “www”, 所以可判定其 internet 取值为 1。
web 中的协议含有“http”“https”“ftp”等关键词。	比如, 温岭市新吉祥铝塑板有限公司在 2008 年 的 web 被错填为“heep://xjxaccp.alu.cn”, 本文 将其修复为“http://xjxaccp.alu.cn”, 由于含有 “http”, 所以可判定其 internet 取值为 1。
web 中含有 IP 地址(4 段数字)。	比如, 乌海市电业局在 2005 年的 web 被错填为 三段数字“10.127.71”, 本文将其修复为 4 段数 字“http://10.127.71.14”, 进而可判定其 internet 取值为 1。
web 中含有中文“网”字且不能是经典门户网站。	比如, 南昌市永胜金属新型门业有限公司的 web 被填为“中国金属门供应网”, 本文可判定其 internet 取值为 1。 但是, 江西汇联实业有限公司的 web 被填为“新 浪网”, 本文暂不判定其 internet 取值为 1。
web 直接为中文“有”字。	

附表 3

附表 3		稳健性检验结果					
	模型 III		模型 IV		模型 V	模型 VI	模型 VII
	<i>internet</i>	标准化偏差	<i>internet</i>	<i>lnfp</i>	<i>new</i>	<i>lnfp_LP</i>	<i>lnfp</i>
<i>internet</i>				0.1309*** (0.0090)	2.2577*** (0.1755)	0.0209*** (0.0037)	
<i>web</i>							0.0886*** (0.0128)
<i>L1.lnmarkup</i>	-0.0955*** (0.0041)	0.5%		0.4586*** (0.0037)	0.6424*** (0.0755)	0.6424*** (0.0755)	0.3849*** (0.0048)
<i>L1.rd</i>					0.3303*** (0.0633)		
<i>market</i>	0.0759*** (0.0024)	0.9%	0.0244*** (0.0038)	0.0690*** (0.0018)	0.5177*** (0.0789)	0.0145*** (0.0014)	0.0611*** (0.0062)
<i>state</i>	0.4611*** (0.0130)	1.3%		-0.0579*** (0.0095)	1.8455*** (0.2506)	-0.0175*** (0.0042)	-0.0959*** (0.0128)
<i>export</i>	0.7814*** (0.0074)	0.7%	0.3439*** (0.0146)	0.1703*** (0.0066)	3.7381*** (0.1591)	0.0349*** (0.0026)	0.1109*** (0.0075)
<i>park</i>	0.1932*** (0.0084)	-0.7%	0.0676*** (0.0182)	0.0465*** (0.0079)	-0.3807** (0.1775)	0.0142*** (0.0028)	0.0572*** (0.0089)
<i>lnage_ind</i>	2.1309*** (0.2106)	1.2%		0.8613*** (0.1657)	-17.6878*** (3.1578)	0.0460 (0.0485)	1.0673*** (0.1531)
<i>ln²age_ind</i>	-0.4556*** (0.0424)			-0.1530*** (0.0331)	3.8338*** (0.6452)	-0.0017 (0.0097)	-0.2048*** (0.0306)
<i>liquid</i>	-0.0032*** (0.0002)	-0.5%		0.0097*** (0.0001)	-0.0193*** (0.0027)	0.0013*** (0.0001)	0.0100*** (0.0002)
<i>debt</i>	-0.0000 (-0.0001)	-0.4%	-0.0007*** (0.0003)	-0.0039*** (0.0001)	-0.0101*** (0.0019)	-0.0009*** (0.0001)	-0.0033*** (0.0004)
<i>diversity</i>	0.6755*** (0.0069)	0.6%	0.2294*** (0.0144)	0.0175*** (0.0061)	0.6685*** (0.1394)	-0.0024 (0.0022)	0.0137** (0.0070)
<i>finance</i>	0.0000 (0.0000)	0.2%		-0.0000*** (0.0000)	0.0000 (0.0001)	-0.0000 (0.0000)	-0.0000*** (0.0000)
<i>subsidy</i>				-0.0003*** (0.0000)	-0.0049 (0.0063)	-0.0001*** (0.0000)	-0.0003*** (0.0001)
<i>L1.internet</i>			2.7095*** (0.0164)				
<i>internet_super_neighbor</i>			2.6761*** (0.1306)				

<i>i.province</i>	Yes		Yes		Yes	Yes	Yes
<i>i.city</i>				Yes			
<i>i.year</i>	Yes		Yes	Yes	Yes	Yes	Yes
<i>i.industry</i>			Yes		Yes	Yes	Yes
<i>i.industry3</i>	Yes			Yes			
lambda				-0.0180** (0.0069)			
PSM 之前处理组 与对照组之差	0.1408*** (0.0036)						
PSM 之后 ATT	0.0789*** (0.0040)						
观测值	880225		77712		87967	79035	77720
R ²	Pseudo: 0.1509		Pseudo: 0.5373		组内: 0.0891	组内: 0.1864	组内: 0.1261

注：圆括号内为稳健标准误；*、**、***分别代表在 10%、5%和 1%的水平下显著。