



```

1 . cd "C:\Users\eliven\Dropbox\ELLW_2026\code"
   C:\Users\eliven\Dropbox\ELLW_2026\code

2 . doedit "02_process_csmar_yearly.do"

3 . do "C:\Users\eliven\Dropbox\ELLW_2026\code\02_process_csmar_yearly.do"

4 . *****
5 . ***# 0. Defining Default Paths
6 .
7 . global path "datasets\yearly"

8 . global rawdata "$path\rawdata"

9 . global data "$path\data"

10 .
11 . *****
12 . *****
13 . ***# 1. importing rawdata
14 . // 1.1 Balance Sheet
15 . import excel "$rawdata\资产负债表140430535\FS_Combas.xlsx", sheet("sheet1") firstrow ca
    > se(lower) clear
    (12 vars, 23,997 obs)

16 . labone, nrow(1) // ssc install labone

17 . drop in 1/2
    (2 observations deleted)

18 . destring _all, replace
    stكد: all characters numeric; replaced as long
    shortname: contains nonnumeric characters; no replace
    accper: contains nonnumeric characters; no replace
    typrep: contains nonnumeric characters; no replace
    a001101000: all characters numeric; replaced as double
    (200 missing values generated)
    a001212000: all characters numeric; replaced as double
    (3 missing values generated)
    a001218000: all characters numeric; replaced as double
    (178 missing values generated)
    a001000000: all characters numeric; replaced as double
    a002101000: all characters numeric; replaced as double
    (5861 missing values generated)
    a002201000: all characters numeric; replaced as double
    (11131 missing values generated)
    a002203000: all characters numeric; replaced as double
    (19873 missing values generated)
    a002000000: all characters numeric; replaced as double

19 .
20 . gen year = year(date(accper, "YMD"))

21 . keep if typrep == "A"
    (0 observations deleted)

22 .
23 . global tofill "a001101000 a001212000 a001218000 a002000000 a002101000 a002201000 a00220
    > 3000"

```

```

24 . foreach var in $tofill{
    2.      replace `var' = 0 if `var' ==.
    3. }
(200 real changes made)
(3 real changes made)
(178 real changes made)
(0 real changes made)
(5,861 real changes made)
(11,131 real changes made)
(19,873 real changes made)

25 . // generate variables
26 . gen at = a001000000

27 . label variable at "asset total"

28 .
29 . gen size = log(a001000000)

30 . label variable size "log(asset total)"

31 .
32 .
33 . gen cash = a001101000 / a001000000

34 . label variable cash "cash scaled by total asset"

35 .
36 . gen ppe = a001212000 / a001000000

37 . label variable ppe "ppe scaled by total asset"

38 .
39 . gen intangible = a001218000 / a001000000

40 . label variable intangible "intangible scaled by total asset"

41 .
42 . gen lev = a002000000 / a001000000

43 . label variable lev "leverage(liab/asset)"

44 .
45 . gen lev_debt = (a002101000 + a002201000 + a002203000) / a001000000

46 . label variable lev_debt "leverage(debt/asset)"

47 .
48 . keep stkcd year at size cash ppe intangible lev lev_debt

49 . compress
    variable year was float now int
    (47,990 bytes saved)

50 . save "$data\BalanceSheet.dta",replace
    file datasets\yearly\data\BalanceSheet.dta saved

51 .
52 . *****
53 . // 1.2 Income Statement

```

```

54 . import excel "$rawdata\利润表140646578\FS_Comins.xlsx", sheet("sheet1") firstrow case(1
    > ower) clear
    (9 vars, 24,445 obs)

55 . labone, nrow(1) // ssc install labone

56 . drop in 1/2
    (2 observations deleted)

57 . destring _all, replace
    stkcd: all characters numeric; replaced as long
    shortname: contains nonnumeric characters; no replace
    accper: contains nonnumeric characters; no replace
    typrep: contains nonnumeric characters; no replace
    b001100000: all characters numeric; replaced as double
    (4 missing values generated)
    b001101000: all characters numeric; replaced as double
    (482 missing values generated)
    b001216000: all characters numeric; replaced as double
    (2255 missing values generated)
    b001300000: all characters numeric; replaced as double
    b002000000: all characters numeric; replaced as double

58 .
59 . gen year = year(date(accper,"YMD"))

60 . keep if typrep == "A"
    (0 observations deleted)

61 .
62 . merge 1:1 stkcd year using "$data\BalanceSheet.dta", nogen

```

Result	Number of obs
Not matched	<b>450</b>
from master	<b>449</b>
from using	<b>1</b>
Matched	<b>23,994</b>

```

63 .
64 . gen roa = b002000000 / at
    (450 missing values generated)

65 . label variable roa "return on assets"

66 .
67 . gen roe = b002000000 / (at - at * lev)
    (450 missing values generated)

68 . label variable roe "return on equity"

69 .
70 . gen rd = b001216000 / at
    (2,604 missing values generated)

71 . label variable rd "R&D scaled by assets"

72 .

```

```

73 . replace b001101000 = b001100000 if b001101000 == . & b001100000 != .
    (478 real changes made)

74 . xtset stkcd year

    Panel variable: stkcd (unbalanced)
    Time variable: year, 2019 to 2023, but with a gap
    Delta: 1 unit

75 .
76 . gen earn = b002000000
    (1 missing value generated)

77 . gen oper_earn = b001300000
    (1 missing value generated)

78 . gen rev = b001101000
    (5 missing values generated)

79 .
80 . global to_lag "earn oper_earn rev"

81 . foreach var in $to_lag{
    2.     forvalues i = 1/5{
    3.         gen l`i'`var' = l`i'`.`var'
    4.     }
    5. }
    (5,588 missing values generated)
    (10,941 missing values generated)
    (15,950 missing values generated)
    (20,463 missing values generated)
    (24,444 missing values generated)
    (5,588 missing values generated)
    (10,941 missing values generated)
    (15,950 missing values generated)
    (20,463 missing values generated)
    (24,444 missing values generated)
    (5,591 missing values generated)
    (10,942 missing values generated)
    (15,951 missing values generated)
    (20,464 missing values generated)
    (24,444 missing values generated)

82 .
83 . egen oper_earn_vol = rowstd(l1oper_earn l2oper_earn l3oper_earn l4oper_earn l5oper_earn)
    (10,940 missing values generated)

84 . label variable oper_earn_vol "sd of past 5 years operating earnings"

85 . gen oper_earn_growth = (oper_earn - l1oper_earn) / l1oper_earn
    (5,588 missing values generated)

86 . label variable oper_earn_growth "operating earnings growth"

87 .
88 . egen earn_vol = rowstd(l1earn l2earn l3earn l4earn l5earn)
    (10,940 missing values generated)

89 . label variable earn_vol "sd of past 5 years earnings"

```

```

90 . gen earn_growth = (earn - l1earn) / l1earn
    (5,588 missing values generated)

91 . label variable earn_growth "earnings growth"

92 .
93 . egen rev_vol = rowstd(l1rev l2rev l3rev l4rev l5rev)
    (10,942 missing values generated)

94 . label variable rev_vol "sd of past 5 years revenues"

95 . gen rev_growth = (rev - l1rev) / l1rev
    (5,594 missing values generated)

96 . label variable rev_growth "revenue growth"

97 .
98 . * For operating earnings:
99 . egen mean_oper_earn = rowmean(l1oper_earn l2oper_earn l3oper_earn l4oper_earn l5oper_earn)
    (5,587 missing values generated)

100 . replace oper_earn_vol = oper_earn_vol / mean_oper_earn
    (13,504 real changes made)

101 . label variable oper_earn_vol "Scaled SD of past 5 years operating earnings"

102 .
103 . * For earnings:
104 . egen mean_earn = rowmean(l1earn l2earn l3earn l4earn l5earn)
    (5,587 missing values generated)

105 . replace earn_vol = earn_vol / mean_earn
    (13,504 real changes made)

106 . label variable earn_vol "Scaled SD of past 5 years earnings"

107 .
108 . * For revenues:
109 . egen mean_rev = rowmean(l1rev l2rev l3rev l4rev l5rev)
    (5,589 missing values generated)

110 . replace rev_vol = rev_vol / mean_rev
    (13,502 real changes made)

111 . label variable rev_vol "Scaled SD of past 5 years revenues"

112 .
113 . keep stkcd year roa roe rd earn_vol earn_growth oper_earn_vol oper_earn_growth rev_vol
    > rev_growth

114 . compress
    variable year was float now int
    (48,888 bytes saved)

115 . save "$data\IncomeStatement.dta",replace
    file datasets\yearly\data\IncomeStatement.dta saved

116 .
117 . *****

```

```

118 . // 1.3 BasicInfo
119 . import excel "$rawdata\上市公司基本信息年度表142226979\STK_LISTEDCOINFOANL.xlsx", sheet
    > ("sheet1") firstrow case(lower) clear
    (11 vars, 62,963 obs)

120 . labone, nrow(1) // ssc install labone

121 . drop in 1/2
    (2 observations deleted)

122 . destring _all,replace
    symbol: all characters numeric; replaced as long
    shortname: contains nonnumeric characters; no replace
    enddate: contains nonnumeric characters; no replace
    industrynamec: contains nonnumeric characters; no replace
    industrycodec: contains nonnumeric characters; no replace
    establishdate: contains nonnumeric characters; no replace
    listingdate: contains nonnumeric characters; no replace
    provincecode: all characters numeric; replaced as long
    (21 missing values generated)
    province: contains nonnumeric characters; no replace
    citycode: all characters numeric; replaced as long
    (109 missing values generated)
    city: contains nonnumeric characters; no replace

123 .
124 . rename symbol stkcd

125 . gen year = year(date(enddate,"YMD"))

126 . drop if year < 2019
    (39,264 observations deleted)

127 .
128 . gen establish_year = year(date(establishdate,"YMD"))
    (7 missing values generated)

129 . label variable establish_year "years of establishment"

130 . gen list_year = year(date(listingdate,"YMD"))

131 . label variable list_year "years of listing"

132 .
133 . gen age_establish = year - establish_year
    (7 missing values generated)

134 . label variable age_establish "years since establishment"

135 .
136 . gen age_list = year - list_year

137 . label variable age_list "years since listing"

138 .
139 . sort stkcd year

140 . global fill2023 "industrynamec industrycodec"

141 . foreach var in $fill2023{
    2.         replace `var' = `var'[_n-1] if year == 2023
    3. }
    (5,354 real changes made)
    (5,354 real changes made)

```

```

142 .
143 . keep stkcd year industrynamec industrycodec establish_year age_establish list_year age_
    > list provincecode province citycode city

144 . order stkcd year industrynamec industrycodec establish_year age_establish list_year age
    > _list provincecode province citycode city

145 .
146 . compress
    variable year was float now int
    variable establish_year was float now int
    variable age_establish was float now byte
    variable list_year was float now int
    variable age_list was float now byte
    variable industrycodec was str13 now str3
    (521,334 bytes saved)

147 . save "$data\BasicInfo.dta",replace
    file datasets\yearly\data\BasicInfo.dta saved

148 .
149 . *****
150 . // 1.4 IndustryClass - Other
151 . import excel "$rawdata\上市公司行业分类年度表161544479\STK_IndustryClassAnl.xlsx", shee
    > t("sheet1") firstrow case(lower) clear
    (15 vars, 104,310 obs)

152 . labone, nrow(1) // ssc install labone

153 . drop in 1/2
    (2 observations deleted)

154 . destring _all,replace
    institutionid: all characters numeric; replaced as long
    enddate: contains nonnumeric characters; no replace
    industryclassificationid: contains nonnumeric characters; no replace
    securityid: all characters numeric; replaced as double
    symbol: all characters numeric; replaced as long
    shortname: contains nonnumeric characters; no replace
    industryclassification: contains nonnumeric characters; no replace
    industrycode1: contains nonnumeric characters; no replace
    industryname1: contains nonnumeric characters; no replace
    industrycode2: contains nonnumeric characters; no replace
    industryname2: contains nonnumeric characters; no replace
    industrycode3: all characters numeric; replaced as long
    (23094 missing values generated)
    industryname3: contains nonnumeric characters; no replace
    industrycode4: all characters numeric; replaced as long
    (81339 missing values generated)
    industryname4: contains nonnumeric characters; no replace

155 .
156 . rename symbol stkcd

157 . gen year = year(date(enddate,"YMD"))

158 .
159 . keep if industryclassificationid == "P0211" | industryclassificationid == "P0218"
    (81,281 observations deleted)

```

```

160 . keep stkcd year industrycode1 industryname1 industrycode2 industryname2 industrycode3 i
    > ndustryname3

161 .
162 . rename industrycode1 swindcode1

163 . label variable swindcode1 "申万一级行业代码"

164 . rename industryname1 swindname1

165 . label variable swindname1 "申万一级行业名称"

166 .
167 . rename industrycode2 swindcode2

168 . label variable swindcode2 "申万二级行业代码"

169 . rename industryname2 swindname2

170 . label variable swindname2 "申万二级行业名称"

171 .
172 . rename industrycode3 swindcode3

173 . label variable swindcode3 "申万三级行业代码"

174 . rename industryname3 swindname3

175 . label variable swindname3 "申万三级行业名称"

176 .
177 . destring _all,replace
    stkcd already numeric; no replace
    swindcode1: all characters numeric; replaced as long
    swindname1: contains nonnumeric characters; no replace
    swindcode2: all characters numeric; replaced as long
    swindname2: contains nonnumeric characters; no replace
    swindcode3 already numeric; no replace
    swindname3: contains nonnumeric characters; no replace
    year already numeric; no replace

178 . compress
    variable year was float now int
    variable swindname1 was str48 now str12
    variable swindname2 was str60 now str24
    (1,703,998 bytes saved)

179 . save "$data\sw_ind.dta",replace
    file datasets\yearly\data\sw_ind.dta saved

180 .
181 .
182 . *****
183 . // 1.5 EquityNature
184 . import excel "$rawdata\中国上市公司股权性质文件140847730\EN_EquityNatureAll.xlsx", shee
    > t("sheet1") firstrow case(lower) clear
    (7 vars, 23,284 obs)

185 . labone, nrow(1) // ssc install labone

```



```

186 . drop in 1/2
    (2 observations deleted)

187 . destring _all,replace
    symbol: all characters numeric; replaced as long
    shortname: contains nonnumeric characters; no replace
    enddate: contains nonnumeric characters; no replace
    largestholderrate: all characters numeric; replaced as double
    toptenholdersrate: all characters numeric; replaced as double
    equitynature: contains nonnumeric characters; no replace
    equitynatureid: contains nonnumeric characters; no replace

188 .
189 . rename symbol stkcd

190 . gen year = year(date(enddate,"YMD"))

191 .
192 . gen soe = regexm(equitynature, "国企")

193 . label variable soe "state owned"

194 .
195 . keep stkcd year largestholderrate toptenholdersrate soe

196 . order stkcd year largestholderrate toptenholdersrate soe

197 . save "$data\EquityNature.dta",replace
    file datasets\yearly\data\EquityNature.dta saved

198 .
199 . *****
200 . // 1.6 Institutional Holdings
201 . import excel "$rawdata\机构持股分类统计表145152956\INI_HolderSystematics.xlsx", sheet("
    > sheet1") firstrow case(lower) clear
    (35 vars, 83,945 obs)

202 . labone, nrow(1) // ssc install labone

203 . drop in 1/2
    (2 observations deleted)

204 . destring _all,replace
    symbol: all characters numeric; replaced as long
    enddate: contains nonnumeric characters; no replace
    fundholdshares: all characters numeric; replaced as double
    (15277 missing values generated)
    fundholdproportion: all characters numeric; replaced as double
    (15277 missing values generated)
    fundholdproportion1: all characters numeric; replaced as double
    (15278 missing values generated)
    qfiiholdshares: all characters numeric; replaced as double
    (72951 missing values generated)
    qfiiholdproportion: all characters numeric; replaced as double
    (72951 missing values generated)
    qfiiholdproportion1: all characters numeric; replaced as double
    (73000 missing values generated)
    brokerholdshares: all characters numeric; replaced as double
    (54434 missing values generated)
    brokerholdproportion: all characters numeric; replaced as double
    (54434 missing values generated)
    brokerholdproportion1: all characters numeric; replaced as double
    (54609 missing values generated)
    insuranceholdshares: all characters numeric; replaced as double
    (75899 missing values generated)
    insuranceholdproportion: all characters numeric; replaced as double
    (75899 missing values generated)
    insuranceholdproportion1: all characters numeric; replaced as double
    (75899 missing values generated)
    securityfundholdshares: all characters numeric; replaced as double
    (72738 missing values generated)
    securityfundholdproportion: all characters numeric; replaced as double

```

```
(72738 missing values generated)
securityfundhproportion1: all characters numeric; replaced as double
(72738 missing values generated)
entrustholdshares: all characters numeric; replaced as double
(77283 missing values generated)
entrustholdproportion: all characters numeric; replaced as double
(77283 missing values generated)
entrustholdproportion1: all characters numeric; replaced as double
(77283 missing values generated)
financeholdshares: all characters numeric; replaced as long
(83817 missing values generated)
financeholdproportion: all characters numeric; replaced as double
(83817 missing values generated)
financeholdproportion1: all characters numeric; replaced as double
(83817 missing values generated)
bankholdshares: all characters numeric; replaced as double
(83005 missing values generated)
bankholdproportion: all characters numeric; replaced as double
(83005 missing values generated)
bankholdproportion1: all characters numeric; replaced as double
(83007 missing values generated)
nonfinanceholdshares: all characters numeric; replaced as double
(74668 missing values generated)
nonfinanceholdproportion: all characters numeric; replaced as double
(74668 missing values generated)
nonfinanceholdproportion1: all characters numeric; replaced as double
(74705 missing values generated)
otherholdshares: all characters numeric; replaced as double
(4058 missing values generated)
otherholdproportion: all characters numeric; replaced as double
(4058 missing values generated)
otherholdproportion1: all characters numeric; replaced as double
(4303 missing values generated)
totalholdshares: all characters numeric; replaced as double
insinvestorprop: all characters numeric; replaced as double
insinvestorprop1: all characters numeric; replaced as double
(245 missing values generated)
```

```
205 .
206 . rename symbol stkcd

207 . gen year = year(date(enddate,"YMD"))

208 . keep if insinvestorprop != .
      (0 observations deleted)

209 . sort stkcd enddate

210 . bys stkcd year: keep if _n == _N
      (55,117 observations deleted)

211 .
212 . keep stkcd year totalholdshares insinvestorprop insinvestorprop1

213 . order stkcd year totalholdshares insinvestorprop insinvestorprop1

214 . save "$data\InstitutionalHoldings.dta",replace
      file datasets\yearly\data\InstitutionalHoldings.dta saved

215 .
```

```

216 . *****
217 . // 1.7 Monthly Return
218 . import excel "$rawdata\月个股回报率文件205559561\TRD_Mnth.xlsx", sheet("sheet1") firstrow
    > ow case(lower) clear
    (13 vars, 302,682 obs)

219 . labone, nrow(1) // ssc install labone

220 . drop in 1/2
    (2 observations deleted)

221 . destring _all, replace
    stkcd: all characters numeric; replaced as long
    trdmnt: contains nonnumeric characters; no replace
    mopnprc: all characters numeric; replaced as double
    clsdt: all characters numeric; replaced as byte
    mclsprc: all characters numeric; replaced as double
    mnshttrd: all characters numeric; replaced as double
    mnvaltrd: all characters numeric; replaced as double
    msmvosd: all characters numeric; replaced as double
    msmvttl: all characters numeric; replaced as double
    ndaytrd: all characters numeric; replaced as byte
    mretwd: all characters numeric; replaced as double
    (1955 missing values generated)
    mretn: all characters numeric; replaced as double
    (1955 missing values generated)
    markettype: all characters numeric; replaced as byte

222 .
223 . gen year = year(date(trdmnt,"YM"))

224 .
225 . gen ln_mret = log(1 + mretwd)
    (1,955 missing values generated)

226 . gsort stkcd year -trdmnt

227 . drop if year == 2024
    (37,469 observations deleted)

228 .
229 . gen exist_return = (mretwd != .)

230 . bys stkcd year: egen count_mret = sum(exist_return)

231 .
232 . // requiring at least 6 months
233 . collapse (sum) ln_mret (sd) mretwd (first) msmvosd msmvttl if count_mret >= 6, by(stkcd
    > year)

234 .
235 . label variable msmvosd "年末流通市值"

236 . label variable msmvttl "年末总市值"

237 .
238 . gen yret = exp(ln_mret) - 1

239 . label variable yret "return of fiscal year"

```

```

240 .
241 . rename mretwd retvol

242 . label variable retvol "volatility of monthly return in fiscal year"

243 .
244 . merge 1:1 stkcd year using "$data\BalanceSheet.dta",nogen

```

Result	Number of obs
Not matched	2,249
from master	97
from using	2,152
Matched	21,843

```

245 .
246 . gen mtb = msmvttl / (at - at * lev) * 1000
      (2,249 missing values generated)

247 . label variable mtb "market to book ratio"

248 .
249 . gen tobinq = (msmvttl + at * lev) / at
      (2,249 missing values generated)

250 . label variable tobinq "Tobin's Q"

251 .
252 . keep stkcd year yret retvol msmvosd msmvttl mtb tobinq

253 . order stkcd year yret retvol msmvosd msmvttl mtb tobinq

254 .
255 . save "$data\MonthlyReturn.dta",replace
      file datasets\yearly\data\MonthlyReturn.dta saved

256 .
257 . *****
258 . // 1.8 Segment
259 . import excel "$rawdata\上市公司子公司情况表21321732\FN_Fn061.xlsx", sheet("sheet1") fir
      > strow case(lower) clear
      (3 vars, 525,083 obs)

260 . labone, nrow(1) // ssc install labone

261 . drop in 1/2
      (2 observations deleted)

262 . destring _all,replace
      stkcd: all characters numeric; replaced as long
      enddate: contains nonnumeric characters; no replace
      fn_fn06101: contains nonnumeric characters; no replace

263 .
264 . gen year = year(date(enddate,"YMD"))

265 . egen grp_seg = group(fn_fn06101)

```

```

266 .
267 . collapse (count) grp_seg, by(stkcd year)

268 . rename grp_seg num_seg

269 . label variable num_seg "number of segments"

270 .
271 . keep stkcd year num_seg

272 . compress
    variable year was float now int
    variable num_seg was long now int
    (75,392 bytes saved)

273 . save "$data\Segments.dta",replace
    file datasets\yearly\data\Segments.dta saved

274 .
275 . *****
276 . // 1.9 AuditorAnalyst
277 . import excel "$rawdata\上市公司基本信息特色指标表213351852\AF_CFEATUREPROFILE.xlsx", sh
    > eet("sheet1") firstrow case(lower) clear
    (10 vars, 23,163 obs)

278 . labone, nrow(1) // ssc install labone

279 . drop in 1/2
    (2 observations deleted)

280 . destring _all,replace
    stkcmec: contains nonnumeric characters; no replace
    stkcd: all characters numeric; replaced as long
    accper: contains nonnumeric characters; no replace
    companysize: all characters numeric; replaced as double
    (172 missing values generated)
    big4: contains nonnumeric characters; no replace
    outside: contains nonnumeric characters; no replace
    anaattention: all characters numeric; replaced as byte
    (10283 missing values generated)
    reportattention: all characters numeric; replaced as int
    (10280 missing values generated)
    companyopacity: contains nonnumeric characters; no replace
    registercapital: all characters numeric; replaced as double

281 .
282 . gen year = year(date(accper,"YMD"))

283 . keep stkcd year big4 outside anaattention reportattention

284 . order stkcd year big4 outside anaattention reportattention

285 .
286 . replace anaattention = 0 if anaattention ==. & big4 != ""
    (10,075 real changes made)

287 . replace reportattention = 0 if reportattention ==. & big4 != ""
    (10,072 real changes made)

288 .

```

```

289 . compress
      variable year was float now int
      variable big4 was str45 now str1
      variable outside was str45 now str1
      (2,084,490 bytes saved)

290 . save "$data\AuditorAnalyst.dta",replace
      file datasets\yearly\data\AuditorAnalyst.dta saved

291 .
292 . *****
293 . // 1.10 Patent
294 . import excel "$rawdata\专利明细情况145138197\PT_LCDETAIL.xlsx", sheet("sheet1") firstrow
      > w case(lower) clear
      (14 vars, 66,145 obs)

295 . labone, nrow(1) // ssc install labone

296 . drop in 1/2
      (2 observations deleted)

297 . destring _all,replace
      symbol: all characters numeric; replaced as long
      enddate: contains nonnumeric characters; no replace
      patentname: contains nonnumeric characters; no replace
      patenttypecode: contains nonnumeric characters; no replace
      patenttype: contains nonnumeric characters; no replace
      declaredate: contains nonnumeric characters; no replace
      applicationnumber: contains nonnumeric characters; no replace
      patentnumber: contains nonnumeric characters; no replace
      applicationdate: contains nonnumeric characters; no replace
      grantdate: contains nonnumeric characters; no replace
      legalstatuscode: contains nonnumeric characters; no replace
      legalstatus: contains nonnumeric characters; no replace
      applicant: contains nonnumeric characters; no replace
      validityperiod: all characters numeric; replaced as double
      (55374 missing values generated)

298 .
299 . rename symbol stkcd

300 . gen year = year(date(enddate,"YMD"))

301 .
302 . gen id_current = (legalstatuscode == "S5101")

303 . gen id_invalid = (legalstatuscode == "S5102" | legalstatuscode == "S5110" | legalstatus
      > code == "S5111" | legalstatuscode == "S5112")

304 . gen id_application = (legalstatuscode == "S5103" | legalstatuscode == "S5104" | legalst
      > atuscode == "S5105" | legalstatuscode == "S5106" | legalstatuscode == "S5107")

305 . gen id_fail = (legalstatuscode == "S5108" | legalstatuscode == "S5109")

306 .
307 . collapse (count) id_current id_invalid id_application id_fail, by(stkcd year)

308 .
309 . rename id_current vaild_patent

```

```

310 . rename id_invalid invalid_patent
311 . rename id_application patent_application
312 . rename id_fail patent_fail
313 .
314 . label variable vaild_patent "number of valid patents"
315 . label variable invalid_patent "number of invalid patents"
316 . label variable patent_application "number of patent applications"
317 . label variable patent_fail "number of patent application fail"
318 .
319 . keep stkcd year vaild_patent invalid_patent patent_application patent_fail
320 . compress
    variable year was float now int
    variable vaild_patent was long now int
    variable invalid_patent was long now int
    variable patent_application was long now int
    variable patent_fail was long now int
    (35,020 bytes saved)
321 . save "$data\Patent.dta",replace
    file datasets\yearly\data\Patent.dta saved
322 .
323 . *****
324 . // 1.11 Accounting quality
325 . use "$data\acc_quality.dta", replace
326 . keep stkcd year aemjones NewsCount ifRestate
327 . duplicates drop

    Duplicates in terms of all variables

    (0 observations are duplicates)
328 . save "$data\Accounting_quality.dta",replace
    file datasets\yearly\data\Accounting_quality.dta saved
329 .
330 . *****
331 . **# 2. Merge together
332 . use "$data\BasicInfo.dta",clear
333 . gen master = 1
334 .
335 . merge 1:1 stkcd year using "$data\sw_ind.dta",nogen

```

Result	Number of obs
Not matched	<b>828</b>
from master	<b>749</b>
from using	<b>79</b>
Matched	<b>22,948</b>

336 . merge 1:1 stkcd year using "\$data\Segments.dta",nogen

Result	Number of obs
Not matched	4,952
from master	4,940
from using	12
Matched	18,836

337 . merge 1:1 stkcd year using "\$data\BalanceSheet.dta",nogen

Result	Number of obs
Not matched	367
from master	80
from using	287
Matched	23,708

338 . merge 1:1 stkcd year using "\$data\IncomeStatement.dta",nogen

Result	Number of obs
Not matched	467
from master	49
from using	418
Matched	24,026

339 . merge 1:1 stkcd year using "\$data\InstitutionalHoldings.dta",nogen  
(variable **year** was **int**, now **float** to accommodate using data's values)

Result	Number of obs
Not matched	6,523
from master	1,095
from using	5,428
Matched	23,398

340 . merge 1:1 stkcd year using "\$data\Patent.dta",nogen

Result	Number of obs
Not matched	26,419
from master	26,419
from using	0
Matched	3,502

341 . merge 1:1 stkcd year using "\$data\MonthlyReturn.dta",nogen

Result	Number of obs
Not matched	5,847
from master	5,838
from using	9
Matched	24,083



```
342 . merge 1:1 stkcd year using "$data\EquityNature.dta",nogen
```

Result	Number of obs
Not matched	<b>6,648</b>
from master	<b>6,648</b>
from using	<b>0</b>
Matched	<b>23,282</b>

```
343 . merge 1:1 stkcd year using "$data\AuditorAnalyst.dta",nogen
```

Result	Number of obs
Not matched	<b>6,825</b>
from master	<b>6,797</b>
from using	<b>28</b>
Matched	<b>23,133</b>

```
344 . merge 1:1 stkcd year using "$data\Accounting_quality.dta",nogen
```

Result	Number of obs
Not matched	<b>13,983</b>
from master	<b>6,799</b>
from using	<b>7,184</b>
Matched	<b>23,159</b>

```
345 .
```

```
346 . keep if (stkcd > 0 & stkcd < 200000) | (stkcd >= 300000 & stkcd < 400000) | (stkcd >= 600000 & stkcd < 700000) // 沪深A股主板(包括中小板002、SH科创板688、SZ创业板300)
(2,052 observations deleted)
```

```
347 . keep if master == 1
(12,321 observations deleted)
```

```
348 . drop master
```

```
349 .
```

```
350 . sort stkcd year
```

```
351 .
```

```
352 . global tofill "vaild_patent invalid_patent patent_application patent_fail"
```

```
353 . foreach var in $tofill{
    2.         replace `var' = 0 if `var' == .
    3. }
(19,503 real changes made)
(19,503 real changes made)
(19,503 real changes made)
(19,503 real changes made)
```

```
354 .
```

```
355 . compress
variable year was float now int
variable soe was float now byte
variable NewsCount was float now int
variable ifRestate was float now byte
(227,690 bytes saved)
```

```
356 . save "$path\merged_yearly_controls.dta",replace  
    file datasets\yearly\merged_yearly_controls.dta saved
```

```
357 .  
    end of do-file
```

```
358 .
```