جلسه دوازدهم:

آنالیزهای چند متغیره

خوشه بندی (Cluster Analysis)

ابتدا به بررسی رویه های آنالیز مربوطه در SAS می‌پردازیم و سپس انجام این آنالیز را در نرم افزارهای STATGRAPHICS و SAS JMP دنبال می کنیم.

در کل این روش بدین صورت کار می‌کند که ابتدا ماتریس تشابه یا تفاوت روی صفات و یا افراد (چون عموما افراد را طبقه بندی میکنیم بایستی ماتریس مذکور برای افراد تهیه گردد و اگر برای صفات تهیه شود بایستی قبل از آن داده ها استاندارد شوند زیرا صفات واحد دارند و قابل قیاس و طبقه بندی نیستند البته استاندارد سازی برای افراد هم پیشنهاد می شود ولی برای صفات الزامی است) ایجاد می شود که برای صفات کمی معیاری مثل مجذور فاصله اقلیدسی بسیار مناسب است. برای صفات کیفی (مثل مارکرهای مولکولی یا نمره دهی‌ها) چهار معیار اصلی وجود دارد که حال بر مبنای هر یک از این چهار معیار یا حتی بیشتر چند دندروگرام حاصل می‌شود هر دندروگرام یک ضریب کوفنوتیک دارد که نشان دهنده میزان همبستگی معیار آن دندورگرام با سایر معیارهای دندروگرام های دیگر است و دندروگرامی برای صفات کیفی معتبرتر است که ضریب کوفنوتیک آن بالاتر باشد یعنی معیاری که با آن ماتریس تشابه یا تفاوت دندروگرام حساب شده بالاترین همبستگی را با سایر معیارها داشته باشد (کار ضریب کوفنوتیک بدست آوردن همبستگی بین معیارهای ایجاد کننده ماتریس تشابه یا تفاوت است). پس برخلاف صفات کمی که معیار فاصله اقلیدسی و یا مجذور فاصله اقلیدسی بهترین است در صفات کیفی معیاری بهترین است که بیشترین همبستگی را با سایر معیارها داشته باشد. در نهایت ماتریس های تشابه و تفاوت با الگوریتم‌هایی (متدهایی) تبدیل به دندروگرام می شوند که مهمترین این متدها برای صفات کمی متد ward و برای کیفی متد UPGMA است.

نکته: اگر دندروگرام حالت پلکانی (chaining effect) پیدا کند متد تبدیل ماتریس به دندروگرام مناسب نبوده و بایستی عوض شود که این حالت عموما در صفات کیفی دیده می‌شود.

در نهایت برای رسم خط گروهبندی دندروگرام از متد CCC-plot و آزمون F-bill استفاده می‌شود هر چند که علم موضوعی در تعداد خوشه بندی و توجه به این نکته که جایی که خوشه ها حین تفکیک بیشترین فاصله را داشته باشند تعداد گروه مناسب حاصل می‌شود و می‌توان خط را رسم نمود. پس زدن خط دندردگرام ترکیبی از آزمون‌های ریاضی و استدلال‌های تئوری است. این خط در SAS JMP و XLSTAT بر مبنای آزمون CCC-plot زده می شود

و نیازی به انجام مجدد آن نیست. با این حال در زیر رویه های مورد نیاز برای رسم دندروگرام و تعیین تعداد گروه (CCC-plot) ذکر شده است. همچنین در کلاس، رسم دندروگرام در نرم افزارهای SAS JMP و STATGRAPHICS آموزش داده شد. برنامه زیر برای داده های کمی است و در نرم افزارهای دیگر هم تاکید آموزشی بر داده های کمی است. برای داده های کیفی علاوه بر نرم افزارهای مذکور می‌توان از نرم افزار NTSYS نیز استفاده نمود.

یک اصل مهم: تجزیه خوشه‌ای از جمله روش‌های چند متغیره گروهبندی است که از صد درصد اطلاعات داده‌ها برای گروهبندی استفاده می‌کند فلذا برای گروهبندی بهتر افراد (تیمارها)، باید از وارد کردن صفاتی که با یکدیگر داری همراستایی هستند یا توانایی کمی در تفکیک تیمارها دارند خودداری نمود. عدم رعایت این اصل یکی از عمومی‌ترین مشکلات دانشجویان در گروهبندی تیمارها توسط آنالیز خوشه بندی است و ناچارا مجبور به انجام سایر روش‌ها مانند بای پلات و ... می‌شوند.

تفسیر: برای تفسیر کلاستر بعد از گروهبندی افراد و اشاره به تعداد گروه بایستی صفاتی که به صورت معنی دار در هر گروه ماکزیمم و مینیمم هست شناسایی و ذکر شود (آزمون F-bill در این راستا کمک کننده است) و در واقع جدول معنی داری صفات بین گروه ها در کنار دندروگرام آورده شود. علاوه بر این برای بررسی تنوع درون گروهی (در صورتی که هر گروه تعداد اعضای زیادی داشته باشد و بخواهیم از هر گروه چند عضو انتخاب کنیم) می‌توان نمودار صورت فلکی را رسم کرد که درواقع دندروگرام به شکل یک صورت فلکی نمایش داده می‌شود. این نمودار تنها با SAS JMP رسم می‌شود (این نرم افزار در تجزیه خوشه ای بالاترین کیفیت را دارد). این نمودار علاوه بر این تعداد گرایی از تعداد گروه واقعی هم بدست می‌دهد زیرا که هر شاخه فلکی در واقع یک خوشه واقعی است و این نمودار هم می‌تواند در رسم خط کمک کننده باشد.

```
data;
input treat$ LacA Ar Tya Gly MI;
cards;
S0    13214118    70    90    80    5834799
S2    17965379.8  90    144319.8    29435.8     313166.4
S4    26218690.2  3467  803275.4    1924016.2   123424
S6    339957650.8 25689 3612981.6   9401425     391727.2
P0    24870169    50    70    60    2129729.6
P2    55579699.4  123   283756.6    1668344.8   549880.4
```

```
P4    82635129.8  9563  531238.2    3004102.6   86
P6    229791623.6 89053 7623419.2   8074714     77
A0    753970      30    40    30    3780711.2
A2    14338200.2  40    134020.2    427112.2    209512
A4    39447019.8  1054  416646.6    2492712.6   75
A6    143600278   96345 2505462.8   7943401     85
;
proc cluster method=ward std ccc;
var LacA Ar Tya Gly MI;
id treat;
run;
proc tree horizontal;
id treat;
run;

proc plot;
plot _ccc_ * _ncl_=_ncl_ /haxis=0 to 16 by 2;
run; quit;
```

آزمون F-bill: در این آزمون برای اطمینان از صحیح بودن گروهبندی و همچنین یافتن معنی داری سهم هر یک از صفات در هر خوشه، هر خوشه (گروه) یک تیمار در نظر گرفته می‌شود و اعضای داخل آن گروه تکرار در نظر گرفته می‌شود اینک یک طرح کاملا تصادفی نامتعادل (زیرا که امکانش کم است که تمامی گروه ها به یک اندازه عضو داشته باشند) اجرا و معنی داری صفات بین تیمارها (گروه‌ها) بررسی می‌گردد. اجرای طرح کاملا تصادفی نامتعادل در جلسه طرح های پایه توضیح داده شده است.

در قسمت آخر برنامه یک رویه پلات قرار دارد که در واقع نتیجه حاصل از تجزیه ccc را روی محور x ی که همان تعداد کلاستر است (nc1) و از صفر تا ۱۶ مختصات بندی شده رسم می‌کند. حال از روی این پلات میتوانیم تعداد خوشه را متوجه بشیم که البته جوابی که میخواهیم متوجه بشیم در خود رویه کلاستر در SAS 9.4 در یک نمودار ارائه داده می شود و در واقع این رویه (proc plot) برای این ورژن نیاز نیست و برای ورژن های قبلی SAS نیاز است.

نکته: به طور کلی در اکثر رویه های رگرسیونی و چند متغیره SAS 9.4 با همان رویه آنالیز اصلی پلات ها را هم می دهد و دیگر مثل ورژن های قبلی نیازی به نوشتن جداگانه رویه های پلات برای بسیاری از تجزیه ها نیست.

موفق باشید