روابط بین متغیرها

برای پیدا کردن روابط بین متغیرهای پارامترهای آماری همچون کوواریانس، همبستگی و رگرسیون موجود هستند. زمانی که دو متغیر تصادفی باشند برای پیدا کردن ارتباط بینشان از همبستگی استفاده می‌شود مثلا پیدا کردن ارتباط بین صفات. همبستگی برای متغیرهای کمی همبستگی پیرسون و برای کیفی اسپیرمن می‌باشد. برنامه هر یک شرح داده می‌شود. فرض همبسیتگی پیرسون نرمال بودن داده‌هاست و باید از تکرارها میانگین گرفته شود. البته اگر بعد از میانگین-گیری تعداد داده‌ها از ۱۵ کمتر می‌شود و نرمالیتی داده‌ها از بین می‌رود بهتر است از میانگین‌گیری صرف نظر کنید.

نکته: در تعداد بالای داده همبستگی‌های پیرسون به طور کاذب معنی دار می‌شوند و بهتر است در این حالت به همبستگی های حتی معنی دار ولی زیر 0.4 اعتنایی نگردد و در تعداد داده پایین (زیر ۳۰ عدد) همبستگی‌ها به سختی معنی دار می شوند.

زمانی که بخواهیم رابطه یک متغیر تصادفی را با یک یا چند متغیر ثابت بسنجیم از رگرسیون استفاده می‌کنیم. رگرسیون علاوه بر نوع ارتباط روند تغییرات را نیز نشان می‌دهد و این کار را با ارائه یک مدل رگرسیونی انجام میدهد که می توان کفایت مدل را هم از طریق پارامترها و تجزیه واریانس رگرسیونی بررسی کرد.

مثلا ارتباط بین سطوح تیماری با یک یا چند صفت. اگر ارتباط با یک صفت سنجیده شود رگرسیون ساده یا تک متغیره نام دارد و اگر با چند صفت سنجیده شود چند متغیره نام دارد. حال هر یک از این حالات (تک متغیره یا چند متغیره) می توانند روابطی خطی یا غیر خطی از خود نشان دهند. در این جلسه با برازش رگرسیونی تک متغیره در حالات خطی و غیر خطی در نرم افزار اکسل آشنا می‌شویم. که در کلاس توضیح داده می‌شود. روند رسم و برازش از طریق سربرگ insert قسمت charts نمودار scatter plot و رویه trending می‌باشد. همچنین رگرسیون خطی ساده و چند متغیره در SAS نیز بررسی می شود.

همبستگی

داده نمونه (دو متغیر آخر کیفی هستند):

solubility/30	solubility/60	WAC/30	WAC/60	turbidity	soluble/insoluble
72.5	74.6	6.1946	6.5793	2	1
71.5	74.4	6.4826	6.587	1	1
74.8	76.9	6.4591	7.4886	0	1
75.5	77.36	6.956	7.4726	0	1
80.3	81.3	3.0293	2.6576	1	2
81.63	82.3	3.5356	3.3026	1	2
82.3	83.5	3.5493	3.5223	2	4
84.8	85.2	3.551	3.495	1	2
68.06	70.3	6.3903	6.4466	3	4
21.4	24.6	35.438	35.8	1	1
24.36	28.4	24.0569	23.8963	0	1
72.9	73.9	6.7803	7.5136	3	3
72.6	73.9	6.85	6.464	2	1
73.69	77.5	7.139	7.6133	1	2
75.6	77.4	7.0457	7.3466	1	1
80.9	81.9	2.974	2.6912	3	2
80.5	82.9	3.3406	3.0116	2	3
81.3	82.9	3.698	3.236	2	4
83.6	84.6	3.685	3.236	1	1
68.4	69.9	5.9943	5.855	3	2
20.3	23.5	34.5696	35.2	0	0
23.6	26.4	24.3147	24.6057	0	0
72.7	74.25	6.4874	7.0464	2	2
72.05	74.15	6.6663	6.5255	3	2
74.2	77.2	6.799	7.5509	2	1
75.55	77.38	7.0008	7.4096	2	1
80.6	81.6	3.0016	2.6744	3	2
81.065	82.6	3.4381	3.1571	3	4
81.8	83.2	3.6236	3.3791	1	1

برنامه همبستگی پیرسون:

```
data;
input solubility30 solubility60 WAC30 WAC60;
cards;

;
proc corr;
var solubility30 solubility60 WAC30 WAC60;
 run; quit;
```

برنامه همبستگی اسپیرمن:

```
data;
input turbidity solubleinsoluble;
cards;

;
proc corr spearman;
var turbidity solubleinsoluble;
 run; quit;
```

رگرسیون خطی

در بحث رگرسیون خطی ابتدا رگرسیون خطی ساده را بررسی می‌کنیم و مطابق سایر برنامه نویسی ها در اینجا مدل ما همان مدل رگرسیونی است. با این تفاوت که همانطور که در طرح ها خطای آخر در مدل نوشته نمی شد اینجا هم عرض از مبدا و ضریب رگرسیون نوشته نمی شود یعنی مدل y=a+bx به صورت y=x نوشته می شود. حال اگر گزاره noint بعد از مدل بیاید مدل بدون عرض از مبدا برآورد می شود یعنی در واقع مدل رگرسیونی یک مدل رگرسیونی از مبدا مختصات است. زمانی این مدل مفید است که از لحاظ منطقی وجود عرض از مبدا باعث می شود مقدار yدر حالاتی غیر منطقی برآورد شود. مثلا در جایی x=0 است و در مدل که قرارش بدیم y تنها برابر عرض از مبدا است و حالا اگر عرض از مبدا منفی باشد y ما که مثلا عملکرد بوده منفی برآورد شده در حالی که عملکرد منفی بی معناست. یا جایی که ضریب تبیین با noint کردن بالا برود می بایستی مدل از مبدا مختصات برازش داده شود به شرطی که حضور عرض از مبدا معنی دار نباشد.

در ابتدا جدول تجزیه واریانس را داریم که مدل همان رگرسیون است با درجه آزادی ۱ و خطا همان انحراف از رگرسیون است با درجه آزادی n-2 و کل هم که همان کل است با درجه آزادی n-1 . حال اگر رگرسیون معنی دار

شود یعنی مدل خطی است و در غیر اینصورت مدل غیر خطی است. در قسمت بعدی بایستی به R^2 تعدیل شده توجه کرد زیرا که ضریب تبیین ساده تحت تاثیر تعداد x قرار می گیرد و دچار اریبی می‌شود (در مدلهای ساده که یک x داریم برایش مشکلی پیش نمی آید و در چند متغیره حتما باید با تعدیل شده کار کرد). ضریب تغییرات هم همانند قبل نشانه ای از میزان خطای کار است. در قسمت بعدی ضرایب عرض از مبدا و رگرسیون آزمون می شوند که معنی داری آنها یعنی این ضرایب مناسب هستندو در رابطه با ضریب رگرسیون اگر در مدل حاضر (مدل خطی) معنی دار شود یعنی فرض H_0 به معنی عدم وجود رابطه خطی رد می شود به عبارتی یعنی رابطه خطی وجود دارد و اگر ضریب رگرسیون معنی دار نشود یعنی این مدل که یک مدل خطی هست مناسب نبوده و رابطه خطی وجود ندارد. در قسمت بعدی آنالیز مانده ها را داریم. اختلاف هر x از $\hat{Y}$ یک مانده را تشکیل می دهد که هر چقدر این اختلاف بیشتر باشد یعنی مدل مناسب نیست. پس می توان نام آن ها خطا هم گذاشت. میانگین آنها برابر صفر است و توزیع این مانده ها روی $\hat{Y}$ ها ملاک است (ردیف اول، شکل اول باید نقاط دو طرف خط به تقریبا به یک اندازه و متقارن باشند و اشکال قیفی و لوزی و.. روی این شکل قابل بررسی است و شکل دوم که بایستی نقاط بین دو خط دارای پراکندگی باشند). توزیع آنها همانطور که در سایر مدل ها بایستی نرمال باشد اینجا هم باید توزیع خطاها نرمال باشد. این در حالتی است که مانده ها توزیع پراکنده داشته باشند. این توزیع نشان می‌دهد خطاها (مانده ها) مستقل از $\hat{Y}$ هستند و در واقع یعنی مدل دارای کفایت است. اما اگر توزیع دارای شکل باشد یعنی مدل خوب نیست. مثلا توزیع قیفی توزیع پوآسن را نشان می‌دهد و برای داشتن یک مدل خطی باید تبدیل داده جذری انجام داد. یا شکل لوزی نشان دهنده توزیع دو جمله است و تبدیل داده زاویه ای لازم داریم و اشکالی منحنی وار نشان دهنده این است که چاره ای جز برازش مدل غیرخطی نداریم. برنامه زیر یک برنامه رگرسیون ساده خطی همراه با عرض از مبدا هست:

```
data;
input x y;
cards;
1       75.99
2       73.4
3       77.9
4       78.8
5       82.3
6       83.03
7       84.7
8       86.6
9       72.5
;
proc reg;
model y=x;
 run;
```

حال اگر بخواهیم مدلی چند متغیره و خطی را برازش دهیم از برنامه زیر استفاده می کنیم. خب مدل آن مثلا برای برازش سه متغیر ثابت (سه تیمار یا حتی رابطه بین عملکرد یا اجزای عملکرد که در این حالت اجزای عملکرد بر حسب علم موضوعی می توانند ثابت فرض شوند) برای یک متغیر تصادفی برابر $y=a+bx_1+bx_2+bx_3$ است. که طبق آنچه گفته شد مدل برابر y=x1+x2+x3 خواهد شد که که میتوان اینگونه نوشت: y=x1-x3 یا y=x1 x2 x3 که حالت دوم برای زمانی است که اسم صفات را نوشته باشیم. درجه آزادی رگرسیون در اینجا برابر تعداد متغیر ثابت و درجه آزادی انحراف از رگرسیون مثلا برای مثال مذکور برابر n-3-1 است (در واقع از تفریق درجه آزادی رگرسیون از کل هم می‌توان به آن رسید). کل هم مانند سابق است. تفاسیر مانند سابق است فقط در قسمت آزمون ضرایب رگرسیون هر متغیری که ضریب رگرسیونش معنی دار نباشد را می توان از مدل حذف کرد.

```
data;
input y x1-x3;
cards;
25      75.99 6.1946       0.851
23      73.4  6.4826       0.623
27      77.9  6.4591       0.455
29      78.8  6.956 0.464
34      82.3  3.0293       0.777
35      83.03 3.5356       0.469
36      84.7  3.5493       0.341
37      86.6  3.551 0.347
22      72.5  6.3903       0.882
;
proc reg;
model y=x1-x3;
 run;
```

موفق باشید