جلسه یازدهم

رگرسیون غیرخطی

معروفترین انواع رگرسیون غیرخطی شامل چندجمله‌ای های متعامد، نمایی، لگاریتمی، پروبیت و تابع لجستیک است. که در اینجا به بررسی سه مورد اخیر می‌پردازیم. این نوع از روابط رگرسیونی هم همانند خطی دارای انواع ساده (یک متغیر x در مقابل یک متغیر y) و چندمتغیره (چند x در مقابل یک y) است.

رگرسیون چند جمله ای های متعامد

```
data;
input x x1 y;
x2=x**2;
x3=x1**2;
x4=x**x1;
cards;
1      4      6.3
3      6      11.1
5      8      20
7      10     24
9      7      21
11     5      18
13     3      16
15     2      14
17     1      12
;
proc reg;
model y=x x2;
model y=x x1 x2 x3 x4;
model y=x x3;
model y=x x2 x3;
run;
```

ساده‌ترین مدل چند جمله ای یک چند جمله ای درجه دوم است که معادله آن می تواند به صورت $y=a+b_1X_1+b_{11}X_1^2$ باشد که یک درجه دو ساده است و یا می تواند به صورت یک چند جمله ای درجه دو چند متغیره باشد معادله آن اقسام گوناگونی خواهد داشت که کاملترین آن به صورت $y=a+b_1X_1+b_2X_2+b_{11}X_1^2+b_{22}X_2^2+b_{12}X_1X_2$ است که میتواند با حذف عوامل غیر معنی دار به سایر اقسام مانند $y=a+b_1X_1+b_{22}X_2^2$ و یا $y=a+b_1X_1 +b_{11}X_1^2+b_{22}X_2^2$ که تمامی این مدل ها به ترتیبی که ذکر شد در قسمت مدل

وارد شده است. در نهایت با بررسی معنی‌داری اجزا مدل، میزان ضریب تبیین تصحیح شده و تست باقی مانده ها می توان بهترین مدل غیرخطی (حال ساده یا یکی از انواع چند متغیره) را گزینش نمود.

رگرسیون نمایی

مدل این رگرسیون به صورت $Y=ae^{b1x}$ است که اگر به صورت $Y=e^{a+b1x}$ نوشته شود به آن رگرسیون رشد می گویند و تابع آن (شکل نمودار) متفاوت است. اما معادله اول در حقیقت همان معادله رگرسیون نمایی است. با لگاریتم گیری طبیعی (بر پایه عدد نپر) از طرفین معادله می توان آن را خطی کرد و مدل رگرسیون آن را با رویه reg بررسی کرد این حرف بدین معناست که می‌توان از تبدیل داده لگاریتمی بر پایه ۱۰ و یا عدد نپر روی داده‌ها انجام داد و سپس آن را به شکل خطی برازش داد. ولی در صورتی که بخواهیم مستقیما مدل را بررسی کنیم بایستی از رویه nlin استفاده نماییم. برنامه اول مستقیما مدل نمایی ساده را بررسی می کند و برنامه دوم مدل خطی ساده بر روی داده های تغییر داده شده را بررسی می کند.

نکته اول: در این رویه بر خلاف رویه های glm و reg بایستی ضرایب عرض از مبدا و رگرسیون در نوشتن مدل لحاظ شوند.

نکته دوم: در رویه nlin بایستی مقادیر ضرایب عرض از مبدا و رگرسیون به صورت تقریبی تعریف شوند که این امر یکی از ضعف‌های این برنامه به حساب می‌آید زیرا که حدس زدن ضرایب امری بسیار مشکل است. در مثال زیر از مقادیر رایجی که در این برنامه استفاده می‌شود استفاده شده است.

برنامه اول:

```
data a;
input x y;
cards;
200 22
250 27
300 43
350 60
400 95
450 135
;
proc nlin;
parms a=5 b=0.008;
model y=a*exp(b*x);
run;
proc gplot;
plot y*x;
run;
```

برنامه دوم:

```
data a;
input x y;
x2=log(x);
cards;
200 22
250 27
300 43
350 60
400 95
450 135
;
proc reg;
model y=x2;
run;
```

رگرسیون لگاریتمی

در کل به معادلاتی که با لگاریتم گیری بر پایه ده خطی شوند لگاریتمی می‌گویند که نمونه اصلی آن معادله ساده $y=ab^x$ است. بهتر است برای این مدلسازیها به جای رویه مستقیم nlin از همان رویه reg استفاده شود و در مثال قبل اگر به رویه nlin اشاره شد تنها جهت آشنایی بود و در برنامه زیر با تبدیل داده و سپس رویه reg مدل لگاریتمی خطی شده و مدلسازی انجام شده است.

نکته: در بحث مدلسازی‌های غیرخطی (بجز چند جمله‌ایها و مدلهای تکه‌ای) شاید استفاده از نرم افزارهای SAS، JMP، SPSS و Minitap کارآمدتر باشد زیرا که هم به صورت منوبار است و هم میتواند بدون در نظر گیری مقادیر رگرسیون و عرض از مبدا مستقیما هر نوع مدلسازی را بر داده ها اعمال کرد.

```
data a;
input x y;
x2=log10(x);
cards;
200 22
250 27
300 43
350 60
400 95
450 135
;
proc reg;
model y=x2;
run;
```

نکته آخر: بایستی بدانید اولا که در مدلسازی های رگرسیونی نرمالیتی اشتباهات داده ها تنها برای مدلهای خطی مطرح است و مدلهای غیرخطی الزامی به این شرط برایشان نیست. ثانیا در رگرسیون برخلاف همبستگی میانگین گیری روی تکرارها توصیه نمی‌شود زیرا که هر چقدر تعدد نقاط بالاتر باشد مدلسازی قوت می گیرد.

موفق باشید