داده‌های طولی (longitudinal data)

به داده‌هایی که به طور مکرر و در طی زمان برای هر یک از واحدهای آزمایشی جمع‌آوری شده اند، داده‌های طولی می‌گوییم. (برای هر واحد آزمایشی بیش از یک مشاهده داریم.)

وابستگی در این داده‌ها:
1- وابستگی درون گروهی : وابستگی به علت واحد آزمایشی
2- وابستگی سریالی : وابستگی به علت زمان

1- مشاهدات مربوط به یک واحد آزمایشی به یکدیگر وابسته اند، چرا که تحت تأثیر خصوصیات همان واحد آزمایشی می‌باشند.

2- چون مشاهدات در طی زمان جمع‌آوری شده اند، مشاهده در یک زمان، به مشاهدات زمان‌های قبل و بعد خود وابسته است. بعنوان مثال فشار خون فرد i در زمان t به فشار خون همان فرد در زمان t-1 وابسته است.

داده‌های پانلی (panel data)

به داده‌هایی که به طور مکرر برای واحدهای آزمایشی جمع‌آوری شده اند، بطوری که برای هر واحد آزمایشی بیش از یک مشاهده در دسترس است، داده پانلی می‌گویند.

وابستگی در این داده‌ها: * وابستگی درون گروهی

* چون این داده‌ها در طی زمان اندازه‌گیری نشده اند، وابستگی سریالی نداریم.

نکته: اسم داده‌های پانلی و طولی گاهاً به جای یکدیگر استفاده می‌شوند و بهتر است همیشه برای تشخیص، نوع جمع‌آوری داده‌ها و نوع داده‌ها را مورد بررسی قرار دهیم.

مثال :

داده های طولی : تعداد ۱۰۰ دانش آموز از یک مدرسه انتخاب شده اند ، نمرات درس ریاضی آن ها در طی سه نوبت می پرسیم.

در این مثال ۱۰۰ واحد آزمایشی داریم که برای هر کدام تعداد 3 مشاهده در دسترس است.

$i = 1, 2, \dots, 100$

$t = 1, 2, 3$

$$
\begin{array}{llll}
i=1 : & x_{11} & x_{12} & x_{13} \\
i=2 : & x_{21} & x_{22} & x_{23} \\
\vdots & & & \\
i=100 : & x_{100\,1} & x_{100\,2} & x_{100\,3}
\end{array}
$$

← x_{11} و x_{12} و x_{13} چون مربوط به یک گروه هستند ، به یکدیگر وابسته اند . (وابستگی درون گروهی)

← هم چنین x_{12} (مشاهده دومین اندازه گیری) تحت تاثیر x_{11} (مشاهده اولین اندازه گیری) است که همان وابستگی پویایی است.

داده های پانلی : ۱۰۰ مدرسه را بصورت تصادفی انتخاب می کنیم ، از هر مدرسه 3 نفر انتخاب می کنیم . نمره درس ریاضی آنها را می پرسیم.

در این مثال نیز ۱۰۰ واحد آزمایشی داریم که برای هر کدام تعداد 3 مشاهده در دسترس است.

$i = 1, 2, \dots, 100$

$t = 1, 2, 3$

$$
\begin{array}{llll}
i=1 : & x_{11} & x_{12} & x_{13} \\
\vdots & & & \\
i=100 : & x_{100\,1} & x_{100\,2} & x_{100\,3}
\end{array}
$$

← در این حالت وابستگی درون گروهی داریم . چون مشاهدات x_{i1} و x_{i2} و x_{i3} ، هر سه از مدرسه i (واحد آزمایشی i) انتخاب شده اند ، هم وابسته اند . در اصل نمره ریاضی ۳ نفر در یک مدرسه هم وابسته است ولی نمره ریاضی یک نفر در مدرسه i به نمره ریاضی یک نفر در مدرسه j وابسته نیست .

* حالتی را در نظر بگیرید که ماهیت داده‌ها فقط وابستگی درون گروهی را شامل می‌شود (و یا ما به انتخاب خودمان فقط این نوع وابستگی را در نظر می‌گیریم.) *

← در مدل‌هایی که فعلاً می‌گوییم، تنها همین نوع وابستگی را در نظر می‌گیریم.

◀ فرض کنید برای متغیر پاسخ y_{it} ، تنها یک متغیر توضیحی x_{it} داریم. یک مدل رگرسیونی ساده بصورت زیر خواهد بود:

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}$$

$i = 1, 2, \dots, n$ → تعداد واحدهای آزمایشی

$t = 1, 2, \dots, T$ → تعداد تکرار برای هر واحد آزمایشی

نکته: لزوماً نباید تعداد تکرار برای هر واحد آزمایشی، برای همه واحدها یکسان و برابر T باشد، در برخی بررسی‌ها ممکن است برای هر فرد تعداد T_i تکرار داشته باشیم، یعنی: $t = 1, 2, \dots, T_i$.

در این مدل رگرسیونی از میان فرضیات مدل رگرسیونی فرض ثبات واریانس و فرض ناهمبستگی خطاها برقرار نمی‌باشند.

* در درس مدل خطی داشتیم که اگر فرض ثبات واریانس برقرار نباشد، برآوردگرها هم چنان ناریب می‌مانند ولی دیگر BLUE نیستند و می‌توانیم برآوردگر دیگری پیدا کنیم که واریانس کمتری داشته باشند.

* چون مشاهدات وابستگی درون گروهی دارند، یعنی به عنوان مثال دو مشاهده y_{it} ، $y_{it'}$ مشاهدات مربوط به جزء i ام در زمان t و t' می‌باشند، بنابراین، کواریانس بین آنها صفر نمی‌شود و از این رو فرض ناهمبستگی خطاها برقرار نمی‌باشد:

$$Cov(y_{it}, y_{it'}) \neq 0$$

حال تمام مشاهدات مربوط به جزء i را در بردار $\underline{y}_i$ قرار داده و فرم برداری مدل بالا را در نظر می‌گیریم:

$$\underline{Y}_i = \beta_0 \underline{1} + \beta_1 \underline{X}_i + \underline{U}_i$$

در این حالت چون تمام مشاهدات مربوط به هر بنود (هر بردار) جدا آورده می شود ، $\underline{Y}_i$ ها ناهمبسته هستند و این جا فرض ناهمبستگی برقرار می باشد.

* در واقع مشاهدات جفت از هم ربطی ندارند و بردار $\underline{Y}_i$ و $\underline{Y}_j$ مستقل می شوند.

در این حالت فرم ماتریس واریانس کوواریانس را در نظر می گیریم :

$$Var(\underline{U}_i) = Var(\underline{Y}_i) = \begin{bmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) & \cdots & Cov(Y_{i1}, Y_{iT}) \\ & Var(Y_{i2}) & \cdots & Cov(Y_{i2}, Y_{iT}) \\ & & \ddots & \\ & & & Var(Y_{iT}) \end{bmatrix} =$$

در حالت کلی می توانند $t=1,\dots,T_i$ باشند.

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1T} \\ & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2T} \\ & & & \ddots & \\ & & & & \sigma_T^2 \end{bmatrix}$$

در این حالت (با در نظر گرفتن این ماتریس بصورت قطری) ، بتعداد $\frac{T(T+1)}{2}$ مجهول داریم که برای برآورد این مجهولات محاسبات بسیاری نیاز است و کار مشکلی می باشد.

* ساختار ماتریس را در حالت اینگونه فرض کردیم که تمام واریانس Y_{it} ها با هم برابر نمی باشند و همچنین تمام کوواریانس ها نیز با هم برابر نمی باشند.

بنابراین سعی می کنیم از مدل های استفاده کنیم که ساختار مشخص تری دارند. در اصل ما می توانیم هر نوع وابستگی در مدل بالا وارد کنیم ولی در مدل ها با ساختار ماتریس واریانس - کوواریانس مشخص تر، این وابستگی ها را خاص تر می کنیم.

مدل‌های با اثرات آمیخته (Mixed effect Models)

۱ـ حالت اول.

α_i را به عنوان اثر تصادفی وارد مدل می‌کنیم. در اصل همان گونه که گفتیم فعلا در نظر می‌گیریم که تنها وابستگی درون گروهی داریم (یعنی وابستگی به خاطر واحد آزمایشی) و این وابستگی را از طریق α_i وارد مدل می‌کنیم.

$$y_{it} = \beta_0 + \beta_1 x_{it} + \underbrace{\alpha_i + \epsilon_{it}}_{u_{it}} \longrightarrow \begin{cases} \alpha_i \overset{iid}{\sim} N(0, \sigma_\alpha^2) \\ \text{ind} \\ \epsilon_{it} \overset{iid}{\sim} N(0, \sigma_\epsilon^2) \end{cases}$$

$u_{it} \sim N(0, \sigma^2)$

همان گونه که قبلاً گفتیم فرض ناهمبستگی خطاها را نداریم و اینجا نمی‌توانیم بگوییم ind هستند و iid نداریم.

بنابراین خطاها ناهمبسته می‌شود و فرض ناهمبستگی برقرار می‌باشد.

حال اگر بخواهیم فرم ماتریس واریانس کوواریانس را مشخص کنیم، داریم:

$$Var(y_{it}) = Var(\alpha_i + \epsilon_{it}) = Var(\alpha_i) + Var(\epsilon_{it}) = \sigma_\alpha^2 + \sigma_\epsilon^2$$

$$Cov(y_{it}, y_{it'}) = Cov(\alpha_i + \epsilon_{it}, \alpha_i + \epsilon_{it'}) = Var(\alpha_i) + \underbrace{\cancel{Cov(\epsilon_{it}, \epsilon_{it'})}}_{0} =$$

چون وابستگی پایی را در نظر نگرفتیم.

بنابراین با توجه به اینکه فرم برداری این مدل بصورت زیر است، ماتریس واریانس کوواریانس بصورت زیر می‌شود:

$$\underline{y}_i = \beta_0 \underline{1} + \alpha_i \underline{1} + \beta_1 \underline{x}_i + \underline{\epsilon}_i$$

$$Var(\underline{y}_i) = \begin{bmatrix} \sigma_\alpha^2 + \sigma_\epsilon^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ & \sigma_\alpha^2 + \sigma_\epsilon^2 & \cdots & \sigma_\alpha^2 \\ & & \ddots & \\ & & & \sigma_\alpha^2 + \sigma_\epsilon^2 \end{bmatrix}$$