各类线性模型

2021年9月11日 17:13

• 方差分析与回归的关系：

- **方差分析与线性回归的转换：**
  - 之前已经知道，t 检验和线性回归是完全对应的。先看最简单的单因素方差分析，其实就是让处理效应（或分组因素，即所谓的单因素，对应于回归中的单个自变量）理解成多分类变量（3个处理组，理解为自变量X的3种取值），据此可以建立线性回归模型。但注意到，自变量不是连续性变量，而是分类变量（有时候为有序分类变量，大多时候为无序分类变量），有两种处理方式：一种是进行哑变量化，第二种是采用最优尺度回归（比哑变量法更有优势的处理技巧）。同样的，在多因素方差分析中，将各个因素转换成各个自变量，并进行最优尺度变换，可以建立线性回归。好像有这么个说法，方差分析其实就是线性回归的特例（所有自变量都是分类变量，并且都进行哑变量化）,Wikipedia中有这么一句话：ANOVA is considered to be a special case of linear regression which in turn is a special case of the general linear model. 另外，对于R2这种指标，在ANOVA和线性回归中是等价的。
  - 回归在实际中的应用：
    1、回归应该关注“因果关系”，但是很多数据分析中忽略了这个内在逻辑。有的人把“果”当初因变量，把“因”当初自变量，这样随意建模时存在问题的，一般表现为建模效果不理想。根源在于，很多自变量共同作用导致了因变量，但是因变量和其他自变量共同作用难以导致某个自变量。因此有些大佬建模时会慎重考虑因果关系。
    2、因果互相影响的情况。现实中，存在因变量对自变量也有影响的情况，比如“住院费越多，疗效越好；病人知道这个规律后，为了达到更好的疗效，而增加住院费”。这种情况可以采用**两阶段最小二乘回归。**
    3、回归分析中普通最小二乘法有LINE要求，在违背这些情况时，有不同的应对策略。当自变量类型不满足要求时，可以采用最优尺度回归；当方差不齐时，可以考虑加权最小二乘回归；当自变量之间存在共线性时，可以考虑岭回归或LASSO回归。

- **协变量与分层变量：**
  - 含协变量的方差分析（协方差分析），需与多因素方差分析进行区分（把影响观测值的其他变量当作协变量而不是“因素”，因为这个协变量是连续性变量，无法当成因素处理）。这种情况需先利用回归的方法（将“因素”和“协变量”一起作为自变量，以观测值作为结局变量，进行线性回归建模）消除组间不平衡的协变量的影响，再对校正后的因变量均数进行处理组间比较的方差分析（校正方法是采用回归模型截距进行校正）。[注意用词：方差分析中的因素对应回归模型中的自变量，方差分析中的变量对应回归模型中的因变量。单因素单变量方差分析对应单自变量单因变量模型；多因素多变量方差分析对应多自变量多因变量模型]。
    - **协变量校正adjusted：**若某变量（次要变量，一般为连续性变量或有序分类变量）与结局变量之间存在线性回归关系的，则该变量常被称为协变量，将其纳入考虑称为 adjusted。协变量校正是先将自变量和协变量一起线性建模，然后用截距值对最终模型（因变量与主要自变量的模型）进行校正。
    - **分层分析stratified：**若某变量（次要变量，一般为有序或无序分类变量）与结局变量之间关系不明确，但是其不同取值可能造成自变量与结局变量之间回归关系发生变化（不存在线性回归关系，但与主要自变量之间存在交互作用），则该变量常被称为分层变量，将其纳入考虑称为 stratified。有时候也会将协变量作为分层变量考虑。分层分析是对于不同层分别进行建模（得到的系数值不同），如果分层变量和主要变量之间的交互作用弱，可以进一步对模型进行合并。
- **协变量与自变量：**
  - 自变量是指研究者主动操纵，而引起因变量发生变化的因素或条件，因此自变量被看作是因变量的原因。
  - 协变量：在实验的设计中，协变量是一个独立变量(解释变量)，不为实验者所操纵，但仍影响响应。同时，

它指与因变量有线性相关并在探讨自变量与因变量关系时通过统计技术加以控制 的变量。常用的协变量包括因变量的前测分数、人口统计学指标以及与因变量明显不同的个人特征等。协变量应该属于控制变量的一种。有些控制变量可以通过实验操作加以控制(如照明、室温等)，也称为无关变量；而另一些控制变量由于受实验设计等因素的限制，只能借助统计技术来加以控制,即成了统计分析中的协变量，因而属于统计概念。

- **交互作用与共线性**：
  - 交互作用是指两个变量共同作用对于结局变量的影响，不等于二者分别作用时的影响的加和（类似于生物学中的协同作用和拮抗作用）。对于交互作用明显的，可以采用分层的策略，分别建模，也可以添加交互项进行建模。
  - 共线性是指两个或多个变量之间本来就存在线性相关关系（实际自由度小于表面上的自由度）。对于共线性明显的，可以先进行变量的筛选。

广义线性模型 GLM：

- **线性回归**的基本如下述公式，本质上是想通过观察 x，然后以一个简单的线性函数 h(x) 来预测 y：

$$y = h(x) = w^T x$$

  - dependent variable y是预测目标，也称 response variable。这里有一个容易混淆的点，实际上 y 可以表达三种含义（建模用的是分布，观察到的是采样，预测的是期望）。Linear Regression 的 y 服从高斯分布，具体取值是实数，在线性回归中我们关注的是分布。但是在线性回归中线性模型有着非常强的局限，即 response variable y 必须服从高斯分布；主要的局限是拟合目标 y 的 scale 是一个实数 $(-\infty,+\infty)$。具体来说有以下这俩个问题，所以这时我们使用 Generalized Linear Model 来克服这俩个问题：
    - y 的取值范围和一些常见问题不匹配。例如 count（游客人数统计恒为正）以及 binary（某个二分类问题）。
    - y 的方差是常数 constant。有些问题上方差可能依赖 y 的均值，例如我预测目标值越大方差也越大（预测越不精确）。

  - observed outcome：是我们的 label，有时用 t 区分表示；这是真正观察到的结果，只是一个值。
  - expected outcome；$y=E[y|x]=h(x)$ 表示模型的预测；注意 y 实际上服从一个分布，但预测结果是整个分布的均值 μ，只是一个值。
  - independent variable x。这是我们的特征，可以包含很多维度，一个特征也称为一个 predictor。
  - hypothesis h(x)：线性模型的假设非常简单，即 $h(x)=wTxh(x)=wTx$ inner product of weight vector and feature vector（权重向量和特征向量的内积），被称为 linear predictor。这是就是线性模型，GLM 也是基与此的推广。深入来看，各个维度特征（predictor） xjxj 通过系数 wjwj 线性加和，这一过程将信息进行了整合；而不同的 weight（coefficient） 反映了相关特征不同的贡献程度 。
  -

- 经典的Logistic模型就是其中一例GLM。GLM是普通线性模型的扩展形式，由于普通线性回归的因变量必须服从正态分布，而实际问题中经常会遇到分类问题或计数问题的建模，GLM采用连接函数（Link Function），将因变量的分布进行了扩展，使得因变量只要服从指数分布族即可（如正态分布，二项分布，泊松分布，多项分布等）。**GLM 可以分解为 Random Component、System Component 和 Link Function 三个部分。**
  - Random Component：
    - An exponential family model for the response，这里是指 response variable 必须服从某一 exponential family distribution 指数族分布，即 y|x,w~ExponentialFamily(η) y|x,w~ExponentialFamily(η)，η 指 exponential family 的 natural parameter 自然参数。例如 linear regression 服从 Gaussian 高斯分布，logistic regression 服从 Bernoulli 伯努利分布。指数族还有很多分布如多项分布、拉普拉斯分布、泊松分布等等。
    - 另外，这也可以被称为 Error Structure（它取决于因变量的分布）：error distribution model for the response。对于 Gaussian 的 residual 残差 $\epsilon=y-h(x)$ 服从高斯分布 $N(0,\sigma)$ 是很直

观；但是，例如 Bernoulli 则没有直接的 error term。可以构造其他的 residual 服从 Binomial。

- System Component
  - 为预测部分，又称 linear predictor，是拟合的关键；因为广义线性模型 GLM 本质上还是线性模型，我们推广的只是 response variable y的分布，模型最终学习的目标还是 linear predictor w^Tx中的 weight vector。注意，GLM 的一个较强的假设是 $\eta=w^Tx$，即 y 相关的 exponential family 的 natural parameter η等于 linear predictor。
- Link Function
  - A link function connects the mean of the response to the linear predictor。它为连接变化函数，用于将指数分布族转化成正态分布，或者说，对预测结果进行非线性映射（建立 linear predictor与 label 之间的变换关系），是LM成长为GLM的关键环节。
  - 通过上述的 Random Component 和 Systematic Component，我们已经把 y 和 w^Tx 统一到了 exponential family distribution 中，最终的一步就是通过 link function 建立俩者联系。对任意 exponential family distribution，都存在 link function $g(\mu)=\eta g(\mu)=\eta$，μ 是 分布的均值而 ηη是 natural parameter；例如 Gaussian 的 link function 是 identity（$g(\mu)=\eta$），Bernoulli 的 link function 是 logit（$g(\mu)=\ln\mu 1-\mu=\eta$）。
  - link function 建立了response variable 分布均值（实际就是我们的预测目标） 和 linear predictor 的关系（准确来说，这只在 T(y)=yT(y)=y 条件下成立，但大多数情况都条件都成立，这里不展开说明）。实际上 link function 把原始 yy 的 scale 转换统一到了 linear predictor 的 scale 上。另外，不同分布的 link function 可以通过原始分布公式变换到指数组分布形式来直接推出。
  - link function 是从 label 映射到 linear predictor的过程，link function的反函数称为响应函数 response function。响应函数 把 linear predictor 直接映射到了预测目标 label。较常见的link function如 logit函数（又称log-odds）；较常用的响应函数如 logistic（又称sigmoid，是二分类中的相应函数）和 softmax（是sigmoid的扩展形式，用于多分类问题），这两个都是 logit 的反函数。
  - 因变量为Bernoulli Distribution也就是对二分类问题建模，因变量为Binomial Distribution也就是对多分类问题建模，因变量为Poisson Distribution也就是对计数问题建模（注意区分计数问题和多分类问题）。

- **LM和GLM的对比：**

Linear Regression

- response variable $y \sim N(\eta, \sigma_e^2)$
- link function $\eta = g(\mu) = \mu$, called *identity*
- prediction $h(x) = \mathbf{E}[y|x, w] = \mu = g^{-1}(\eta) = \mu$

Generalized Linear Model

- response variable $y \sim exponential\ family$
- link function $g(\mu)$, eg. logit for Bernoulli
- prediction $h(x) = \mathbf{E}[y|x, w] = \mu = g^{-1}(\eta)$, eg.logistic for Bernoulli

- 它们的预测部分是linear predictor η=wTx是一样的。不过对 response variable 服从分布的假设不一致。GLM 中我们最常用到的是 Logistic Regression
- 无论是 LM 还是 GLM，我们对不同数据 xx 得到的其实是不同的 response variable 的分布，例如 Gaussian 的 response function 是 $g-1(\eta)=\mu g-1(\eta)=\mu$；而 exponential family 根据具体假设的分布，使用相应的 response function （例如 Bernoulli 是 sigmoid）。所以不同分布的 μ 值不同，进而我们预测的结果不同。虽然每一条数据只是预测了一个值，但其实对应的是一个分布。并且数据相似的话对应的分布也相似，那么预测结果也相似。

        - 因为LM是要求响应变量y的分布是正态的，所以很有局限，而GLM的分布类型则广的多，用来拟合的效果会更好。
    - **回归分分析中的交互作用：**

- 线性混合模型 LMM：
    - 首先看一下 Wiki上对混合模型MM的介绍：A mixed model (or more precisely mixed error-component model) is a statistical model containing both fixed effects and random effects.（注意：fixed在这里译为固定，不同于mixed混合）
    - 混合模型擅长于处理纵向数据（重复测量数据）和有缺失的数据，并且往往优于ANOVA等方法。在混合模型中，需要区分两个概念：random effects与 random errors。

      以矩阵定义混合模型，可以写成：

      $$y = X\beta + Z\gamma + \epsilon$$

      y 是观测值的向量，服从多元正态分布，且平均值可以表示为 $E(y) = X\beta$

      β 是固定因子的效应值（与X对应的固定效应参数向量）

      γ 是随机因子的效应值，服从多元正态分布，且平均值为 $E(\gamma) = 0$，它的方差为 $Var(\gamma) = G$

      ϵ 是残差的向量矩阵，它的平均值为 $E(\epsilon) = 0$，它的方差为 $Var(\epsilon) = R$

      X 为**固定效应**自变量的设计矩阵（可包括连续性变量和分类变量，甚至可包含交互项或二次项等），Z 为**随机效应**变量构造的设计矩阵。

    - 切勿将固定效应狭义理解为主要变量，而应该是所有可能的解释变量（如分组变量和时间变量），包括这些变量之间的交互项。而随机效应则是假定的随机效应部分（这部分的意义应当从多水平模型的角度来理解了）。
    - 该模型为固定效应 X\beta 和随机效应 Z\gamma 的混合，且固定效应和随机效应均与响应变量为线性关系，因此称为线性混合模型。注意：当满足球形检验时，重复测量资料的线性混合效应模型可退化为一般线性模型。

      混合模型的假定为 $\gamma \sim N(0, G)$，$\epsilon \sim N(0, R)$，其中 $Cov(\gamma, \epsilon) = 0$，即两者的协方差为0（二者互相独立）。可以给出Henderson's "mixed model equations" (**MME**):

      $$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} X'R^{-1}\boldsymbol{y} \\ Z'R^{-1}\boldsymbol{y} \end{pmatrix}$$

      The solutions to the MME, $\hat{\beta}$ and $\hat{u}$ are best linear unbiased estimates (BLUE) and predictors (BLUP) for β and u（此处的u指的就是γ，有的版本习惯使用 u 来替代 γ 字符），respectively. 拟合混合模型还可以使用 EM 算法。

- 多水平模型 MLM
    - 多水平模型其实和线性混合模型LMM是等价的，只是理解的角度不同而已。MLM是从模型组建的多个水平来理解，关注构建过程；LMM则仅关注模型构建的结果（固定效应部分+随机效应部分）。多水平模型可以分层表述，然后整合成一个公式（即等价于LMM的公式）。
    - 一个两水平的模型例子如下：

一个包含“2个水平1的解释变量（x和z）和1个水平2的解释变量（w）”的两水平模型可以表述为：

$$
\begin{aligned}
y_{ij} &= \beta_{0j} + \alpha_1 x_{1ij} + \beta_{1j} z_{1ij} + e_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01} w_{1j} + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} w_{1j} + u_{1j}
\end{aligned}
$$

其中，$i = 1, 2, ..., N$（N是总样本量），$j = 1, 2, ..., J$（J是水平2的解释变量的w的取值个数，假定w为分类变量）。则 y_{ij} 表示在变量w的第 j 种取值的情况中的第 i 个个体的结局测量值。第1水平方程（第1个等式）中，截距 β_{0j} 带有下标 j，表示其值随 w 的取值变化而变化；系数 β_{1j} 带有下标 j，表示变量 z_{1ij} 对 y_{ij} 的效应随 w 的取值变化而变化；而系数 α_1 不带有下标 j，表示变量 x_{1ij} 对 y_{ij} 的效应不随 w 的取值变化而变化。在两个第2水平方程（第2、3个等式）中，第1水平的回归系数变成了因变量。关于其他参数如e和u的规则，此处跳过（感兴趣的可查阅统计书《高级医学统计学》）。

从概念上来讲，该模型的建立是从顶向下的，先进行第1水平的参数计算（通过枚举 j 来获得 j 组回归系数 β_{0j} 和 β_{1j}）；然后使用估计的回归系数进行第2水平的参数计算，生成多个第2水平的方程。这种计算步骤是传统的计算方法，现在的计算其实是同步进行的。

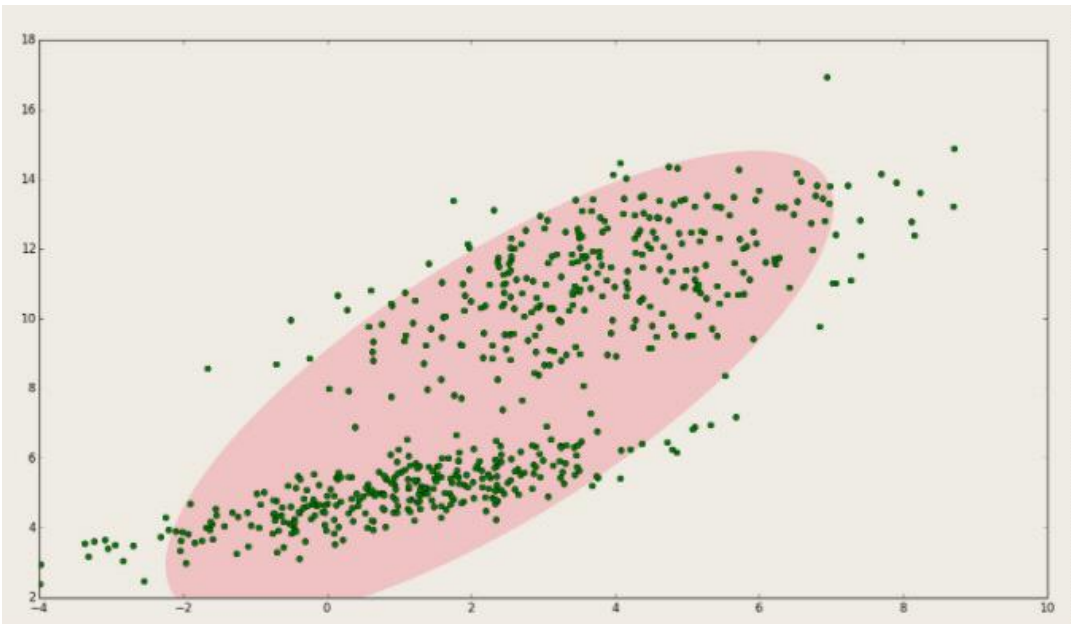
如果将两个第2水平的方程代入到第1水平的方程中，可以得到：

$$
y_{ij} = (\gamma_{00} + \gamma_{01} w_{1j} + \alpha_1 x_{1j} + \gamma_{10} z_{1ij} + \gamma_{11} w_{1j} z_{1ij}) + (u_{0j} + u_{1j} z_{1ij} + e_{ij})
$$

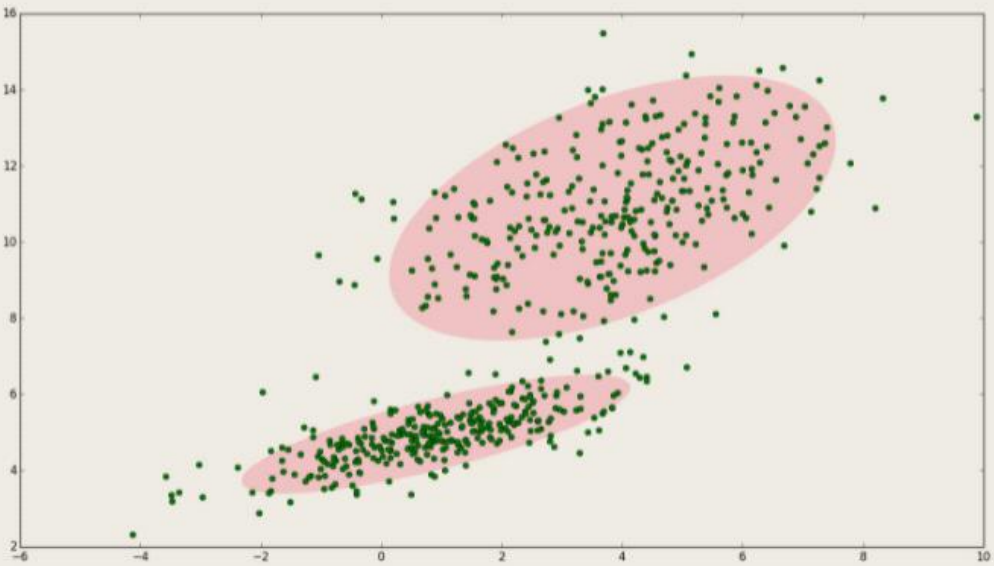
- 这是一个组合模型，该式右边分为两部分，第一个括号部分是各个解释变量及其交互项产生的效应，第二个括号部分是复合残差结构。第一部分便可对应为LMM中提到的固定效应部分，第二部分可对应为LMM中提到的随机效应部分（包括纯粹残差项）。

高斯模型GMM

- 高斯混合模型GMM（Gaussian Mixed Model）指的是多个高斯分布函数的线性组合，理论上GMM可以拟合出任意类型的分布，通常用于解决同一集合下的数据包含多个不同的分布的情况（或者是同一类分布但参数不一样，或者是不同类型的分布，比如正态分布和伯努利分布）。
- 如下面的图：图中的点在我们看来明显分成两个聚类。这两个聚类中的点分别通过两个不同的正态分布随机生成而来。但是如果没有GMM，那么只能用一个的二维高斯分布来描述图1中的数据。图1中的椭圆即为二倍标准差的正态分布椭圆。这显然不太合理，毕竟肉眼一看就觉得应该把它们分成两类。



- 这个时候实际上可以使用GMM了！如下面的图，数据在平面上的空间分布和上面的图一样，这时使用两个二维高斯分布来描述下面图中的数据，分别记为N(μ1,Σ1)和N(μ2,Σ2)。图中的两个椭圆分别是这两个高斯分布的二倍标准差椭圆。可以看到使用两个二维高斯分布来描述图中的数据显然更合理。实际上图中的两个聚类的中的点是通过两个不同的正态分布随机生成而来。如果将两个二维高斯分布N(μ1,Σ1)和N(μ2,Σ2)合成一个二维的分布，那么就可以用合成后的分布来描述图2中的所有点。最直观的方法就是对这两个二维高斯分布做线性组合，用线性组合后的分布来描述整个集合中的数据。这就是高斯混合模型（GMM）。



设有随机变量$\boldsymbol{X}$，则混合高斯模型可以用下式表示：

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

其中$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$称为混合模型中的第$k$个**分量（component）**。如前面图2中的例子，有两个聚类，可以用两个二维高斯分布来表示，那么分量数$K = 2$. π_k是**混合系数（mixture coefficient）**，且满足：

$$\sum_{k=1}^{K} \pi_k = 1$$

$$0 \leq \pi_k \leq 1$$

实际上，可以认为π_k就是每个分量$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$的权重。

- GMM常用于聚类。如果要从 GMM 的分布中随机地取一个点的话，实际上可以分为两步：首先随机地在这 K 个 Component 之中选一个，每个 Component 被选中的概率实际上就是它的系数，选中 Component 之后，再单独地考虑从这个 Component 的分布中选取一个点就可以了—这里已经回到了普通的 Gaussian 分布，转化为已知的问题。可以用EM算法估计GMM参数。
- 另外，还有个概念叫广义矩方法，也简称GMM

广义估计方程 GEE

- 广义估计方程（generalized estimating equation, GEE）用于估计广义线性模型的参数（其中线性模型的结果之间可能存在未知的相关性）。于1986年由Liang和Zeger首次提出，是在广义线性模型和重复测量数据中，运用准似然估计方法估计参数的一种用于分析相关性数据的回归模型。
- 为什么选用广义估计方程：

对于观察值是连续性变量的重复测量资料，一般可以采用**单变量方差分析（ANOVA）**或**多元方差分析（MANOVA）**的方法（最好是连续性变量满足正态性、方差齐性以及各时间点组成的协方差具有球形性）；但**对于离散型重复测量资料**（如变量为二分类变量），**一般采用广义估计方程GEE进行统计分析。**

单变量方差分析（单因素方差分析，ANOVA），就是传统的普通方差分析，将p个时间点类比成p个处理组（这种类比有些拗口），则对应为完全随机设计（总变异=处理变异+误差），可用于单组重复测量资料分析；若本身就存在多个处理组（多组重复测量资料），可再将m个处理组类比成m个区组（拗口的类比方式），则可采用随机区组的单因素方差分析设计（总变异=处理变异+区组变异+误差）。[注意：ANOVA要求p个处理组之间相互独立，因此要求满足球形检验（各时间点的测量值之间互相独立，或者称满足"**独立结构**"）；若不满足球形检验，则需进行校正，否则容易增大第I类错误的风险]

多元方差分析（MANOVA），是将p个时间点看成p维向量，而不是看成一个时间变量的p个水平（不再将其类比为p个处理水平）。由于ANOVA要求的球形检验（各时间点测量值之间互相独立）前提，在很多情况下无法满足，而MANOVA不需要满足球形检验（正好适合处理存在相关性的问题，Hotelling's T2检验的拓展形式）。MANOVA的要求是服从**多元正态分布。**

这部分内容摘自《高级医学统计学》。

- 广义估计方程是在广义线性模型的基础上发展起来的，专门用于处理纵向数据等重复测量资料的统计模型，包括不均衡的纵向数据（纵向数据中研究对象重复测量次数、重复测量间隔时间可能有不同，使得纵向数据不均衡，如队列研究中途研究对象失访。而重复测量方差分析常需满足球形检验）。除了正态分布，GEE利用连接函数将二项分布、Poisson分布、Gamma分布等多种分布的应变量拟合为相应的统计模型，解决了重复测量数据非独立性问题，可得到稳健的参数。
- GEE的特点是：
  （1）只要联接函数g(.)正确， 总观测次数足够大，即使作业相关矩阵Ri(α)指定不完全正确， β的可信区间和模型的其他统计量仍然渐近正确。
  （2）广义估计方程采用准似然估计法估计参数，计算比最大似然法简单，并且对多元分布也没有要求，当样本量较大时， 甚至相关矩阵选择的不合适也对估计的影响不大。 特别是当资料中有缺失值，每个观测对象的观测次数不同，观察时间间隔不同等条件下，都可选用 GEE进行分析。
  （3）广义估计方程应用条件较宽，可适用于多种类型的反应变量，如定量变量、分类变量、等级变量等，同时也可纳入多种类型的自变量，因而在重复测量设计资料统计分析中应用广泛。

GEE 与LR 比较

- Logistic回归方程通常假定所有观察值是相互独立；
- 广义估计方程可输出作业相关矩阵，分析各时间点的相关参数，从而比较不同时间点的差异；
- 广义估计方程还可以探讨各因素的交互作用及对自变量作用的分解，即检验自变量对于不同时间点的影响大小是否相同。

至此，应该理解了重复测量资料为什么不能直接进行Logistic回归建模分析。

- 一般线性模型——局限性：只能拟合因变量服从正态分布的资料，不适用于分类资料。如果说方差分析做的事情本质上和线性回归一样，那么GEE做的事情本质上和广义线性回归是一样的（回归任务推广到分类任务）。
- 广义线性模型——广义估计方程是广义线性模型的延展。其借助线性模型的分析思路解决模型构造、参数估计和模型评价等一系列问题。广义线性模型要求有个均值函数g(μ)，以便把因变量的期望值和线性预测值ηi关联起来。基本结构为：
  $g(\mu) = \eta_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + ... + \beta_j * X_{ij}$
- 联接函数的作用就是对应变量作变换使之符合正态分布，变量变换的类型依应变量的分布不同而不同。优点：用于拟合应变量服从正态分布的模型，拟合服从二项分布、poisson分布、负二项分布的等指数分布族

模型。通过指定不同的联接函数，把指数分布族的众多模型统一到一个模型框架中，具有极大的灵活性。

- 广义估计方程处理重复测量纵向资料的优势：很好地解决了纵向数据的相关性问题，利用了纵向数据中每次测量的结果，大大减少了信息的损失。对于临床试验重复测量资料，广义估计方程能有效地考虑组内相关性，处理有缺失值的资料，可以获得中心效应的参数及其标准误的估计值。以及在考虑了中心效应之后，可以有效估计处理因素有无作用及其作用大小。采用广义估计方程对临床试验重复测量资料进行统计分析，可以使药物疗效评价更为客观。
- 作业相关矩阵是广义估计方程中的一个重要概念，表示的是应变量的各次重复测量值两两之间相关性的大小。作业相关矩阵的形式常有以下几种：只要联接函数正确，总观测次数足够大，作业相关矩阵对参数估计的影响不大。

> （1）**等相关**，又称可交换的相关(exchangeable correla-tion)，或复对称相关(compound symmetry correlation)。假设任意两次观测之间的相关是相等的。这种假设常用于不依时间顺序的重复测量资料。
>
> （2）**相邻相关**，即只有相邻的两次观察值间有相关。
>
> （3）**自相关**(autocorrelation)，即相关与间隔次数有关，相隔次数越长，相关关系越小。
>
> （4）**不确定型相关**(unstructured correlation)，即不预先指定相关的形式，让模型根据资料特征自己估计。
>
> （5）**独立**(independent)，即不相关(uncorrelated)，就是假设应变量之间不相关（多次观察值互相独立）。即**独立结构**或**球形结构**。

广义线性混合模型 GLMM

- 广义线性混合模型GLMM，可以看做是线性混合模型LMM的扩展形式，使得因变量不再要求满足正态分布；也可以看作是GLM的扩展形式，使得可以同时包含固定效应和随机效应。
- LMM模型的一般形式为图一，而GLMM在此基础上做了一些改动。令 linear predictor, \eta, 表示固定效应和随机效应的组合（随机误差不包含在内），即图二

$$y = X\beta + Z\gamma + \epsilon$$

$$\eta = X\beta + Z\gamma$$

- 令g(·)表示link function，用来连接 linear predictor 和 label，h(·)为g(·)的反函数，即response function。则有

$$g(E(y)) = \eta, E(y) = H(\eta) = u, \quad \text{因此：} y = h(\eta) + \epsilon$$

- 带随机效应的Logistic回归中的 probability density function 或简称PDF，和带随机效应的Poisson回归中的probability mass function 或简称PMF）。结果的解读，和GLM中的解读类似，细微的差别仅在于随机效应部分的解读。
- 举个例子，我们认为疗效可能与服药时间相关，但是这个相关并不是简简单单的疗效随着服药时间的变化而改变。更可能的是疗效的随机波动的程度与服药时间有关。比如说，在早上10：00的时候，所有人基本上都处于半饱状态，此时吃药，相同剂量药物效果都差不多。但在中午的时候，有的人还没吃饭， 有的人吃过饭了，有的人喝了酒，结果酒精和药物起了反应，有的人喝了醋，醋又和药物起了另一种反应。显然，中午吃药会导致药物疗效的随机误差非常大。这种疗效的随机误差（而非疗效本身）随着时间的变化而变化，并呈一定分布的情况，必须用广义线性混合模型了。对于固定效应来说，参数的含义是，自变量每变化一个单位，因变量平均变化多少。而对于随机效应而言，参数是服从正态分布的一个随机变量，也就是说对于两个不同的自变量的值，对因变量的影响不一定是相同的。（固定效应仅影响y的平均值；随机效应仅影响y的方差）