残差和预测值

调用广义线性模型或广义相加模型或 COX 模型，对应变量与一组自变量进行回归，获得该应变量的预测值与残差。预测值和/或残差值可用于进一步分析。应变量应是连续性（残差分布符合正态，联系函数=identity）或 0/1 两分类变量。

1. 应变量是连续性：如果自变量有曲线拟合选项，调用广义相加模型（gam），如果没有曲线拟合项，调用广义线性模型（glm）。

2. 应变量是 0/1 两分类的时间依赖的生存状态变量，一般 1 表示事件发生，0 表示生存。需要定义 cox 模型需要的时间变量，同时可以给出需要统一预测的时间。如果自变量有曲线拟合项，调用 R rms 包的 rcs 函数进行曲线拟合。如果给出了统一预测的时间，则输出的预测值为每个个体在该时间时的生成概率，没有残差。如果没有给出统一的预测时间，输出的预测值为每个个体在其各自观察到的时间下的生存概率，残差即观察到的生存状态与预测的生存概率的差。cox 模型预测值可以有如下几种：

（1）根据模型表达式可以计算 linear predictor ("lp"); 如果没有曲线拟合项，该值可以从输出方程中计算。

（2）lp 值可以转换为风险值：risk score = exp(lp)

（3）预测在观察时间时的事件发生数：the expected number of events given the covariates and follow-up time（"expected"）

（4）expected 值可以转换为生存概率：Survival probability = exp(-expected)，本模块输出该值。

3. 应变量是 0/1 两分类非时间依赖的生存状态变量，用 logistic 回归（二项分布，联系函数为 logit），如果自变量有曲线拟合选项，调用广义相加模型（gam），如果没有曲线拟合项，调用广义线性模型（glm）。输出的预测值为 Y=1 的概率，残差为观察到的 Y（0/1）与预测概率的差。

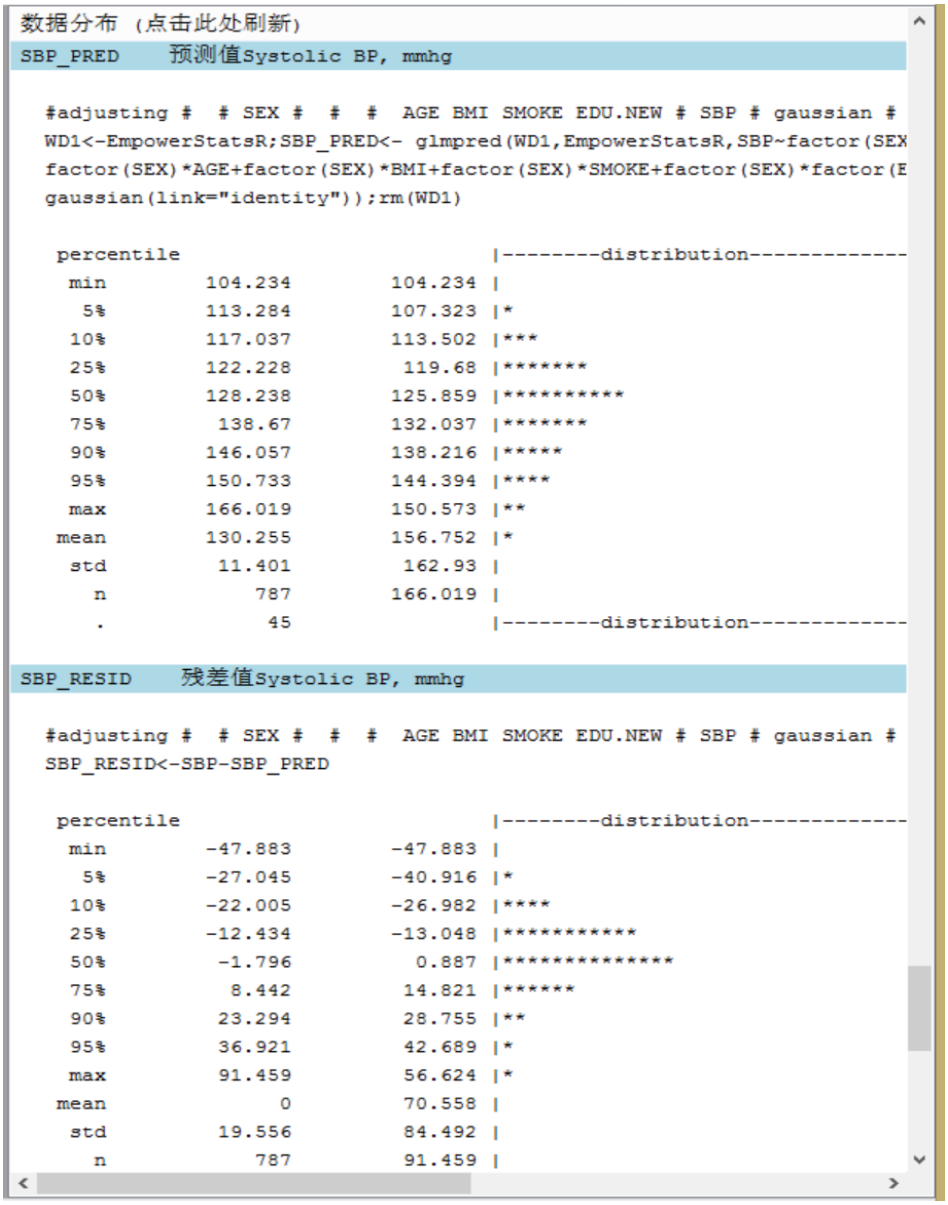
例如，我们要分析血压（收缩压和舒张压）与单核苷酸多态性标记（SNP1） 的关联，我们知道收缩压和舒张压受到年龄、体重指数(BMI)、吸烟和教育程度的影响。我们先调整年龄、体重指数、吸烟和教育程度，计算调整后的收缩压和舒张与收缩压和舒张的残差值，然后把它与 SNP1 进行关联分析。

如果要（用广义相加模型）曲线拟合年龄与体重指数，右击变量名，选曲线拟合（S）即可。如下界面显示：

每个应变量产生两个新变量，新变量名是原变量名后面加"_RESID"与"_PRED"，分别表示残差与预测值。点击后缀名"_RESID"与"_PRED"，可以修改新变量名后缀。

SBP_PRED	预测值Systolic BP, mmhg
SBP_RESID	残差值Systolic BP, mmhg
DBP_PRED	预测值Diastolic BP, mmhg
DBP_RESID	残差值Diastolic BP, mmhg

右击右边变量分布页面中上述新变量名，点击刷新可以查看残差与预测值的分布：



如要预测年龄 60 岁时发生 HBP 的概率，如下界面所示：

此时需要定义时间变量“AGE”与预测时间 60。输出结果将只有 HBP_PRED,即在 60 岁时未发生 HBP（生存）的概率。