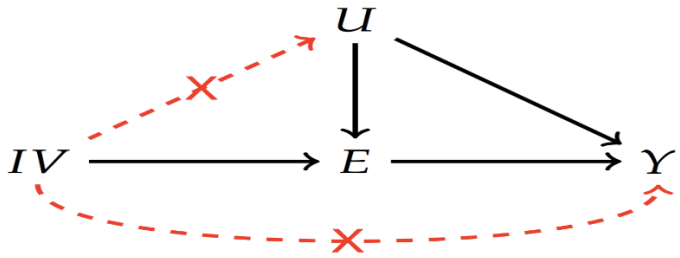
工具变量模型 (Instrument Variables Model)

回归方程拟合暴露因素（E）对结局变量（Y）的作用时可以通过调整已经测量到的混杂因素的作用，来估计 X 对 Y 的独立作用，然而未测量到的混杂因素(U)是无法调整的。因为有未测量到的混杂因素的存在，方程中的残差项与 X 相关，此时 X 被称为内生性变量（endogenous variables），方程中与残差无相关的自变量被称为外生性变量（exogenous variables）。未测量到的混杂因素（内生性）导致的结果可想而知是使得对暴露 E 的效应估计不准。工具变量模型因此而出现。

工具变量的思想是把 E 分解成两部分，一部分与残差相关，这部分是未测量到的混杂因素作用部分；另一部分与残差不相关，这部分是我们需要的。而要实现这种分解，需要借助另一个变量，即工具变量（Instrument variables, IV）。能作为工具变量需要三个条件，如下图所示：



如图示：工具变量须（1）与 E 相关；（2）只通过 E 对 Y 起作用，对 Y 没有其它直接作用；（3）与其它混杂因素（与 E 和 Y 均相关的因素）不相关。这三个条件中，条件（1）可以通过现有数据验证，而条件（2）与（3）是无法验证的，需要研究人员的专业判断。

对条件（2）与（3），使用本模块如果选用 TSLS 方法，系统将自动采用 Aderson-Rubin 敏感性分析，对工具变量通过其它途径（直接作用或通过与未知的混杂因素 U 相关）对 Y 的影响在正负 3%个（Y 的）标准差范围内时，暴露变量 E 的净效应及其 95%可信区间进行估计。如果选用其它工具变量估计方法（如 2SPS，2SRI，GMM），则不能进行敏感性分析。

近年来工具变量模型发展了很多种方法，常用的有：两阶段最小二乘法（2SLS: Two-Stage Least Square）、两阶段预测概率替代(2SPS: Two-Stage Predictor Substitution)，两阶段残差引进(2SRI: Two-Stage Residual Inclusion)，结构均数模型(SMM: Structural Mean Model)及其延申线性模型的广义矩估计(GMM: Generalized Method of Moments) 等。这些方法都有一定的局限。使用本模块，用户可以选择上述四种模型。

两阶段最小二乘法（2SLS: Two-Stage Least Square）：

这是最多知道的两阶段方法，也是最传统的方法。该方法采用两个模型，第一个模型估计工具变量（Z）对暴露变量（X）的效应，计算出每个研究对象 X 的预测值，用该预测值代替 X 的实际观测值，进入下一个模型，用来估计 X 对 Y 的效应。

$$X_i = \alpha_0 + \alpha_z Z_i + \alpha_c C_i + \varepsilon_{1i}; \text{ for } i = 1,2,.....n$$
$$Y_i = \beta_0 + \beta_{IV} \hat{X}_i + \beta_c C_i + \varepsilon_{2i} \text{ for } i = 1,2,.....n$$

当有多个工具变量时，TSLS 估计有偏，这时可采用有限信息最大似然值方法(LIML: Limited Information Maximum Likelihood) 进行估计。使用 LIML 的一个条件是残差的方差要有同质性。如果此条件不满足，可考虑使用 GMM。当结果变量（Y）为两分类变量或暴露（X）与结果（Y）的关系是非直线性关系时，TSLS 方法会做出有偏估计。

两阶段预测概率替代（2SPS）

两阶段预测概率替代是 2SLS 到非线性模型的扩展，用于估计边缘（人口平均）比值比。在第一阶段，使用非线性最小二乘法（NLS）或任何其它有一致性的估计方法来估计工具变量（IV）和暴露（X）之间的关系。然后，用来自第一阶段模型的暴露的预测值替换观察到的暴露状态，进入第二阶段模型。第二阶段模型可以是 logsitic 回归或 Cox 生存分析，对于连续性暴露和结果，2SPS 和 2SLS 结果相似。

两阶段残差引进模型（2SRI）

2SRI 也被称为控制函数估计，是另一种两阶段方法，首先由 Hausman 提出。 2SRI 的一般概念是将第一阶段模型中的误差项（残差）作为附加变量包括在第二阶段模型中。第一和第二阶段的模型可以是线性模型或非线性模型（logistic 模型, Cox 模型）。在线性模型的情况下，2SRI 估计与 2SLS 和 2SPS 估计等价。然而，对于 Logistic 回归模型，由于比值比的非折叠性，2SRI 估计值可能不是因果比值比。

对线性和非线性模型，2SRI 的估计比较一致。 2SRI 优于 2SLS 在于 2SLS 仅在第二阶段模型是线性时是一致的，而 2SRI 不受这种限制。此外，研究显示 2SRI 比 2SPS 估计更精确。

结构均数模型（SMM）与广义矩估计（GMM）

SMM（Structural Mean Models）使用反事实或潜在结果，最初由 Robins 提出用于分析有不依从性的随机试验数据，以估计治疗（暴露）的因果效应。 SMM 是半参数模型，使用工具变量（IV）通过 G 估计来识别和估计因果参数。该方法对暴露的分布不做任何假设。如果使用本身联系函数（identity link），SMM 又被称为相加 SMM（additive SMM）用于连续性的结果变量的分析。使用对数线性模型的乘法 SMM（multiplicative SMM）用于计数型/或两分类结果变量的分析，估计因果风险比（risk ratio）。此外，由 Vansteelandt 和 Goetghebeur 以及 Robins 和 Rotnitzky 开发的逻辑结构均数模型（Logistic Structural Mean Model）用于分析 0/1 二分类元结果变量，估计因果比值比。

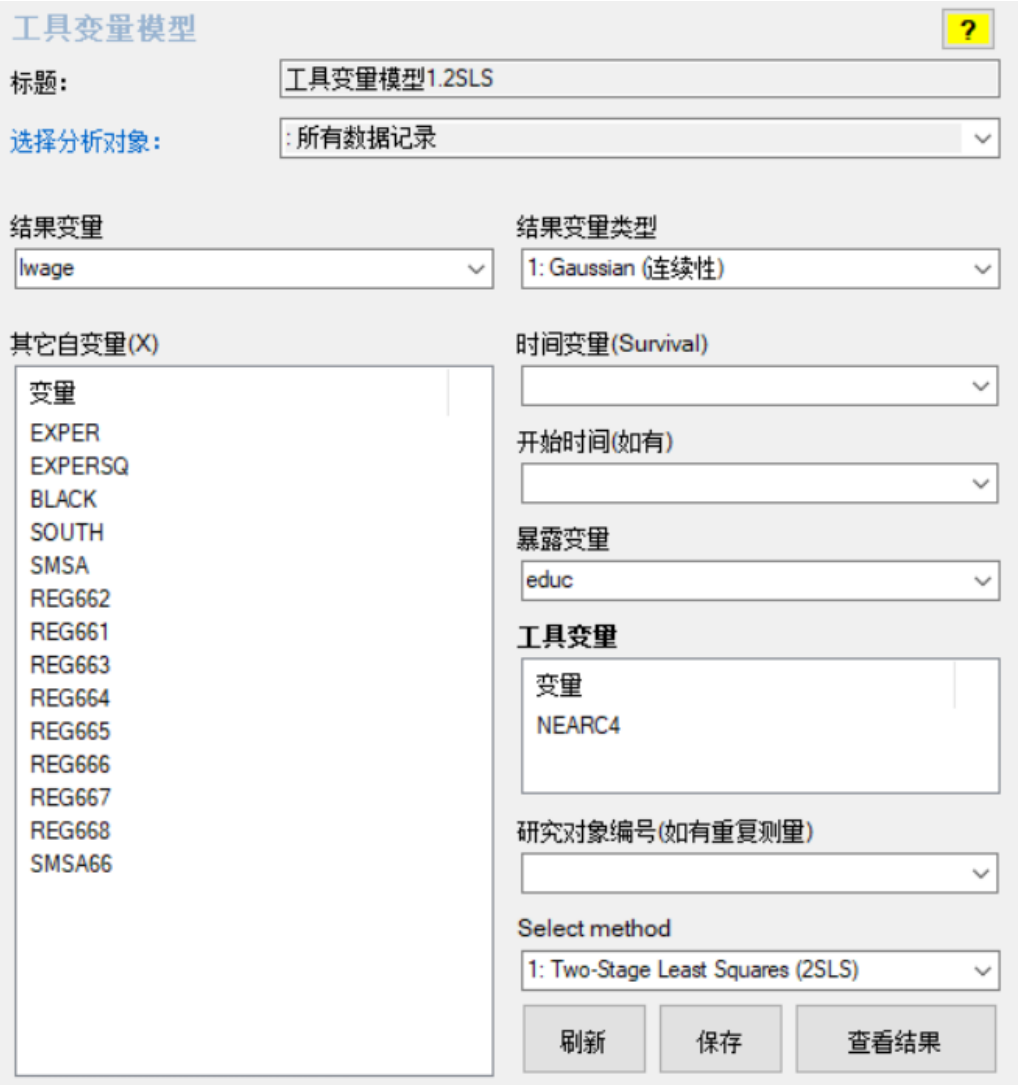
前述的两阶段工具变量方法均要求暴露效应的同质性，即没有效应修饰因子的存在，而这种假设在实际情况下难以验证。SMM 方法的优点在于它没有这种要求。

当工具变量数大于暴露变量数时，采用 GMM (Genearlized Method of Moments) 广义矩量法，该方法由 Hansen 提出，是一类广泛适用的估计方法。在线性模型和单个工具变量的情况下，GMM 等同于 2SLS、相加 SMM 和 LIML。

使用本模块，当选用 Generalized Method of Moments 时，系统采用结构均数模型（如工具变量数>1，自动采用 GMM 估计）。结构均数模型中不包含其它自变量，这时如果用户设置了其它自变量（X），将不参与分析。

例 1：下载练习数据：http://www.empowerstats.com/empowerStats/exdata/card.xls

该数据来自国家青年队列调查数据（Card D 1995. “Using Geographic Variations in College Proximity to Estimate the Return to Schooling.”），分析受教育年数与工资水平的关系，结果变量为 LWAGE，暴露变量为 EDUC，使用工具变量 NEARC4（离大学或学院的距离），其它自变量有 EXPER, EXPERSQ, BLACK, SOUTH, SMSA, REG662, REG661, REG663, REG664, REG665, REG666, REG667, REG668, SMSA66。调用基本统计-回归分析-工具变量模型，输入界面如下：



结果变量类型为连续性，选用 2SLS。点击查看结果，结果输出如下：

```
Instrument variables model (Using R package ivmodel)
Outcome: LWAGE
Exposure: EDUC
Instrument variable: NEARC4
```

Other covariates (exogenous): EXPER; EXPERSQ; BLACK; SOUTH; SMSA; REG662; REG661; REG663; REG664; REG665; REG666; REG667; REG668; SMSA66
Total sample size: 3010

解释： 上面是对输入参数的描述

First stage regression between IV(NEARC4) and exposure(EDUC)

F = 13.26 , df1 = 1 , df2 = 2994 , p-value = 0.000276

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.848524	0.211122	79.804619	0
NEARC4	0.319899	0.087864	3.64085	0.000276
EXPER	-0.412533	0.0337	-12.241485	0
EXPERSQ	0.000869	0.00165	0.526287	0.598728
BLACK	-0.935529	0.093735	-9.980596	0
SOUTH	-0.051613	0.135428	-0.381106	0.703152
SMSA	0.402182	0.104811	3.837208	0.000127
REG662	-0.288907	0.147339	-1.960828	0.049992
REG661	-0.210271	0.202457	-1.038597	0.299076
REG663	-0.23821	0.142636	-1.670059	0.095012
REG664	-0.093089	0.185983	-0.500525	0.616742
REG665	-0.482887	0.188187	-2.565996	0.010336
REG666	-0.513086	0.209635	-2.447517	0.014442
REG667	-0.427089	0.205621	-2.077069	0.03788
REG668	0.31362	0.241674	1.297701	0.19449
SMSA66	0.025481	0.105769	0.240907	0.809644

F statistic is 13.26 , which is > 10, indicating the IV is not weak!

解释：上面是第一阶段模型的输出，满足工具变量的第一个条件是工具变量与暴露相关，这可通过对第一阶段模型进行 F 检验来验证，当 F 值 < 10 时，表示工具变量与暴露的相关性比较弱。F 值的计算方法是比较模型中有工具变量与没有工具变量的两个模型，通过方差分析计算出 F 值。如果只有一个工具变量，其 F 值就是模型输出中该工具变量（NEARC4）对应的 t 值的平方（3.64^2 = 13.26）。

Coefficients of k-Class Estimators:

	k	Estimate	Std. Error	t value	Pr(>|t|)
OLS	0	0.074693	0.003498	21.351022	0
Fuller	0.999666	0.127501	0.052708	2.41899	0.015623
LIML	1	0.131504	0.054964	2.392559	0.016793
TSLS	1	0.131504	0.054964	2.392559	0.016793

Alternative tests for the exposure effect under H0: beta=0.

	Conditional Likelihood Ratio(CLR)	Anderson-Rubin (AR)	AR sensitivity test
F	5.42	5.42	5.42
DF1		1	1
DF2		2994	2994
P value	0.020028	0.020028	0.049504
95% CI lower	0.024804	0.024805	0.000347
95% CI upper	0.284825	0.284824	0.340944

OLS: ordinary least squares
TSLS: two-stage least squares
LIML: limited information maximum Likelihood
Fuller: Fuller's estimator (FULL)
Confidence interval by AR and CLR test are robust even for weak IVs
AR sensitivity analysis with delta range -0.03 to 0.03 (3 percent of SD of outcome)

解释：上面输出第二阶段模型对暴露 EDUC 的效应的估计，有 4 个估计值，OLS 是未采用工具变量的传统估计方法的结果。TSLS（2SLS）即两阶段最小二乘法估计结果，同时输出了 LIML（有限信息最大似然值估计）与 Fuller 估计（Fuller 估计与 LIML 类似）。上述三种工具变量估计出来的效应值相近。当工具变量数大于暴露变量数时，LIML 的结果可靠。

同时进行条件似然比检验与 Anderson-Rubin 检验，当工具变量与暴露变量的相关性不强（F<10）时，这两个检验给出的暴露变量对 Y 的效应的 95%可信区间比较稳定可靠。

Anderson-Rubin 敏感性（AR sensitivity）分析：当工具变量不满足条件 2（对 Y 没有直接作用）与条件 3（与其它既与 Y 又与暴露 E 相关的混杂因素无关）时，即工具变量通过其它途径（不通过暴露 E）对 Y 有作用时，Anderson-Rubin 敏感性分析分析可以判断如果工具变量通过其它途径对 Y 的效应达到 Y 的标准差的 3%时，所估计的暴露变量对 Y 的效应值的可信区间与 P 值。本例结果显示 EDUC 仍然对 LWAGE 有显著效应。

例 2，将例 1 中的工具变量估计方法改选为 Generalized Method of Moments，此时其它自变量设置无效。输出结果如下：

Outcome: LWAGE
Exposure: EDUC
Instrument variable: NEARC4
Instrument variables model using Generalized Method of Moments (R package gmm)

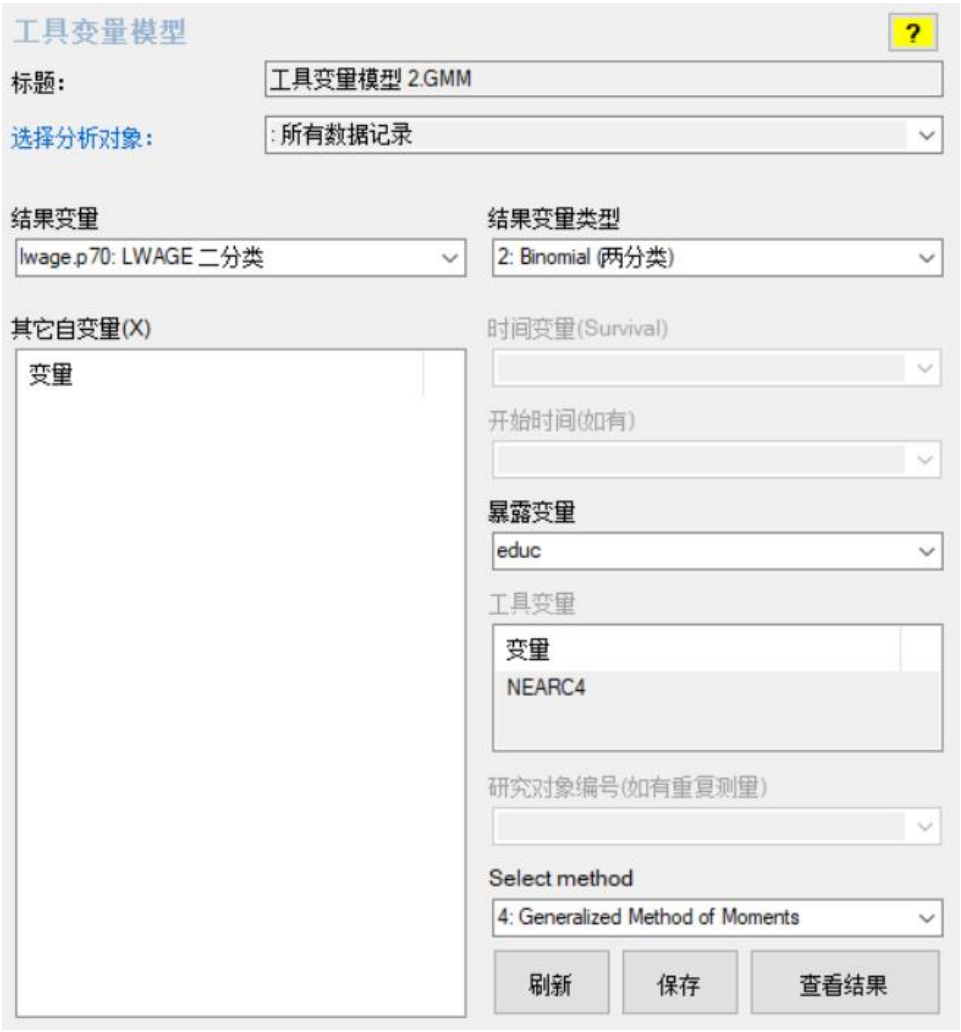
Additive SMM (Structural Mean Model)

	Estimate	Std. Error	t value	Pr(>|t|)	95% CI lower	95% CI upper
(Intercept)	3.767472	0.348746	10.802915	0	3.083942	4.451001
EDUC	0.188063	0.026283	7.155402	0	0.13655	0.239576

解释：LWAGE 为连续性变量，采用相加结构均数模型，输出 EDUC 的效应估计为 0.188063。

注：如果将例 1 中其它自变量除去，采用 2SLS 方法估计，得出的结果与此完全相同。

例 3，首先将 LWAGE 按是否大于 70%分位数转换成一个新的 0/1 两分类变量，新变量名为 LWAGE.P70。将例 2 中结果变量换成两分类变量 LWAGE.P70，采用 Generalized Method of Moments 估计方法。输入界面如下：



输出结果如下：

Outcome: LWAGE 二分类
Exposure: EDUC
Instrument variable: NEARC4
Instrument variables model using Generalized Method of Moments (R package gmm)

Additive SMM (Structural Mean Model)

	Estimate	Std. Error	t value	Pr(>|t|)	95% CI lower	95% CI upper
(Intercept)	-1.53543	0.326793	-4.698474	3e-06	-2.175934	-0.894927
EDUC	0.138282	0.024628	5.614795	0	0.090012	0.186553

解释：相加结构均数模型输出的结果，可以解释为危险差：即 EDUC 每增加一个单位，导致的 LWAGE=1 的风险增加多少。

Multiplicative SMM (Structural Mean Model)

	Estimate	Std. Error	t value	Pr(>|t|)	95% CI lower	95% CI upper
E(Y0)	0.009207	0.002178	4.227936	2.4e-05	0.004939	0.013475
Log causal risk ratio	0.262571	0.01932	13.590582	0	0.224704	0.300437

causal risk ratio	95% CI lower	95% CI upper
1.3003	1.252	1.3504

解释： 相乘结构均数模型输出的结果为危险比，即 EDUC 每增加 1 个单位，LWAGE=1 的风险比是多少。

Alternative Multiplicative SMM (Structural Mean Model)

	Estimate	Std. Error	t value	Pr(>|t|)	95% CI lower	95% CI upper
log(E(Y0))	-11.975789	6.908479	-1.733491	0.083008	-25.516159	1.564581
Log causal risk ratio	0.963116	0.8291	1.161639	0.245382	-0.661891	2.588123

causal risk ratio	95% CI lower	95% CI upper
2.6198	0.5159	13.3048

解释： 替代相乘结构均数模型将 Intercept 替换为 Log(E(Y0)),相当于联系函数为 log(log(Y))，在特定的情况下适用，输出的结果同样为危险比。此处仅供参考。

Logistic SMM (Structural Mean Model)

Model convergence has been reached successfully

	Estimate	Std. Error	t value	Pr(>|t|)	95% CI lower	95% CI upper
Intercept	-3.830306	0.389093	-9.844203	0	-4.592914	-3.067699
X	0.195917	0.028463	6.883213	0	0.140131	0.251704
Z	0.69948	0.475396	1.471364	0.141193	-0.232278	1.631238
X*Z	-0.017637	0.034397	-0.512753	0.608124	-0.085054	0.049779
E(Y0)	0.00023	0.000288	0.799163	0.424196	-0.000334	0.000793
Log causal odds ratio	0.621114	0.133813	4.641656	3e-06	0.358845	0.883383

causal odds ratio	95% CI lower	95% CI upper
1.861	1.4317	2.4191

解释： 逻辑结构均数模型输出的结果为比值比，即 EDUC 每增加 1 个单位，LWAGE=1 的比值比是多少。