

线性回归中的假设检验

慧航

2025年4月



- $$\frac{\hat{\beta}_k}{\text{s.e.}(\hat{\beta}_k)} \sim t(N - K)$$

在统计软件中，通常都会在计算 $\hat{\beta}$ 之后，汇报标准误，并给出在原假设： $H_0 : \beta_k = 0$ 条件下的 $p - value$ 。

回归表格

VARIABLES	(1) y	(2) y	(3) y	(4) y
exer	-1.856*** (0.285)	2.912*** (0.105)	3.187*** (0.0990)	3.059*** (0.0720)
gender				-10.05*** (0.0720)
Constant	77.57*** (0.205)	70.02*** (0.0915)	79.92*** (0.0521)	79.96*** (0.0483)
Observations	1,000	491	509	1,000
R-squared	0.041	0.613	0.671	0.953

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

大样本同方差条件下的分布

- 小样本下的检验仍然依赖于误差项的正态性假设，大样本条件下，可以使用CLT从而可以放弃误差项正态的假设
- 在大样本条件下，我们可以放弃正态性假设，有：

$$\sqrt{N}(\hat{\beta} - \beta) \overset{a}{\sim} \mathcal{N}(0, \sigma^2 [\mathbb{E}(x_i x_i')]^{-1})$$

从而：

$$\sqrt{N}(\hat{\beta}_k - \beta_k) \overset{a}{\sim} \mathcal{N}(0, \sigma^2 [\mathbb{E}(x_i x_i')]_{kk}^{-1})$$

大样本同方差条件下的分布

- 此外，由于：

$$s^2 \xrightarrow{p} \sigma^2, \frac{1}{N} \sum_{i=1}^N x_i x_i' \xrightarrow{p} \mathbb{E}(x_i x_i')$$

从而：

$$s^2 \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right)^{-1} = N s^2 (X'X)^{-1} \xrightarrow{p} \sigma^2 [\mathbb{E}(x_i x_i')]_{kk}^{-1}$$

- 根据Slutsky定理：

$$\frac{\hat{\beta}_k - \beta_k}{\text{s.e.}(\hat{\beta}_k)} = \frac{\sqrt{N}(\hat{\beta}_k - \beta_k)}{\sqrt{N s^2 (X'X)^{-1}_{kk}}} \stackrel{a}{\approx} \mathcal{N}(0, 1)$$

大样本同方差条件下的检验

- 因而对于原假设 $H_0 : \beta_k = b$ ，我们可以使用统计量

$$\frac{\hat{\beta}_k - b}{\text{s.e.}(\hat{\beta}_k)} \stackrel{a}{\sim} \mathcal{N}(0, 1)$$

对原假设进行假设检验。

- 注意到以上的检验统计量与同方差、正态总体条件下的检验统计量是一模一样的，除了其分布变为标准正态分布。
- 实际上：
 - 如果 K 固定，当样本量足够大时， $t(N - K) \xrightarrow{D} \mathcal{N}(0, 1)$ ，此时使用 t 分布或者正态分布并无差异；
 - 另一方面， t 分布的比标准正态分布具有更厚的尾巴，意味着在相同的显著性水平下， $t(N - K)$ 的临界值总是大于标准正态分布的临界值，而在小样本时， t 分布和正态分布都不精确，但是使用 t 分布却有着更小的犯第 I 类错误的概率。
 - 基于以上原因，在实践中，无论样本量大小，我们通常都使用 $t(N - K)$ 分布取临界值。

异方差稳健标准误

- 如果存在异方差，以上计算的标准误都是错误的
- 怀特异方差稳健标准误：

$$\widehat{\mathbb{V}}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{i=1}^N \hat{u}_i^2 x_i x_i' \right) (X'X)^{-1}$$

- 在异方差的条件下，如果我们使用异方差稳健标准误，那么：

$$\frac{\hat{\beta}_k - \beta_k}{\text{s.e.}(\hat{\beta}_k)} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\widehat{\mathbb{V}}(\hat{\beta})_{kk}}} \stackrel{a}{\sim} \mathcal{N}(0, 1)$$

仍然成立。因而假设检验步骤与上述大样本条件下的假设检验步骤相同，唯一的区别在于使用的标准误替换为异方差稳健标准误。

一个实例：OHIE实验中的假设检验

OHIE实验

在美国，Medicaid是针对穷人的健康保险计划。在2008年时，俄勒冈州计划恢复Medicaid中的OHP Standard计划。由于预计申请人数非常多，因而州政府推出了一个按照抽签分配名额的方法。个人一旦被抽中，整个家庭都可以享受该计划。在个人被抽中后，州政府会联系申请人参加计划，然而由于种种原因，并非所有抽中的人最终都参加了该计划。为了检验实际数据中抽签是否仍然是随机的，Finkelstein等人(2012)使用了如下回归：

$$y_i = \beta_0 + \beta_1 \times \text{treatment}_i + u_i$$

其中 treatment_i 为个体 i 是否抽中签（=1为处理组，=0为对照组）， y_i 为一些人口统计指标，比如出生年份、性别、语言等等，如果 $\hat{\beta}_1$ 不显著，那么可以认为抽中和没有抽中标签的两组人具有相似的特征。（ohie_qje_test_one.do）

一个实例：OHIE实验中的假设检验

OHIE实验

变量	(1) 出生年份	(2) 性别	(3) 英语问卷	(4) 本人签署	(5) 首日填报	(6) 有电话号码
处理组	0.253* (0.148)	-0.0172*** (0.00615)	-0.00647* (0.00370)	-0.0440*** (0.00438)	0.00530 (0.00351)	-0.00196 (0.00421)
常数项	1.968*** (0.106)	0.556*** (0.00439)	0.909*** (0.00262)	0.876*** (0.00296)	0.0900*** (0.00247)	0.871*** (0.00298)
Observations	33,823	33,822	33,823	33,823	33,823	33,823
R-squared	0.000	0.000	0.000	0.004	0.000	0.000

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

多重共线性

- 在最小二乘法的假设中，要求 $\sum_{i=1}^N x_i x_i' = X'X$ 可逆，如果该假设不满足，我们称为完美共线性（perfect collinearity）。
 - 如果存在完美共线性，意味着变量之间可以完美的线性表示出，或者存在完美的线性相关性，那么回归方程是不可识别的。
- 然而现实的情况可能是，虽然 $X'X$ 可逆，但是变量之间的相关性非常高，此时，留给每一个变量的变异（variation）可能会太小，从而导致对该系数的推断有失准确。

系数的标准误

- 令 $X = \begin{bmatrix} X_{-*} & x_* \end{bmatrix}_{N \times K}$, 其中 x_* 为一个1维列向量, 可能是我们所关注的变量, 那么

$$X'X = \begin{bmatrix} X'_{-*}X_{-*} & X'_{-*}x_* \\ x'_*X_{-*} & x'_*x_* \end{bmatrix}$$

- 使用分块矩阵求逆公式, 假设 X_{-*} 包含常数项, 有:

$$\begin{aligned} (X'X)^{-1}_{KK} &= \left[x'_*x_* - x'_*X_{-*} (X'_{-*}X_{-*})^{-1} X'_{-*}x_* \right]^{-1} \\ &= \frac{1}{(N-1) s_{x_*}^2 (1 - R_*^2)} \end{aligned}$$

其中 R_*^2 为使用 x_* 对 X_{-*} 做回归所得到的可决系数。

多重共线性的后果

- 多重共线性会导致一些我们不希望看到的后果，比如：
 - 由于回归系数的方差比较大，因而所估计出的系数非常不稳定，精度很差，甚至会出现本来为正，估计出来却是负的情况，或者相反；
 - 同样由于由于回归系数的方差比较大，虽然估计出来的系数比较大，但是由于标准误也非常大，假设检验的结果可能是不显著，此时更容易犯第II类错误。
 - 问题：第I类错误呢？

多重共线性的模拟

多重共线性的后果

	b	多重共线性	N	拒绝率	$\bar{\hat{b}}$	s.d. ($\hat{b}$)	s.e. ($\hat{b}$)	neg
(1)	0	有	20	5.85%	-0.026	0.794	0.769	50.3%
(2)	1	有	20	23.6%	0.979	0.768	0.804	10.3%
(3)	1	无	20	94.65%	0.991	0.243	0.253	0.05%
(4)	1	有	200	85.65%	1.007	0.321	0.327	0.1%

多重共线性的后果

- 多重共线性的存在并没有像异方差问题一样，使得标准误的估计出现问题。
 - 多重共线性导致的问题是在给定样本量的条件下，估计系数的标准差变大了。
 - 由于标准误的估计没有问题，因而犯第I类错误的概率也没有被影响。
- 但是，由于标准误变大了，犯第二类错误的概率却变大了，这意味着，即使某个变量有影响，由于多重共线性的存在，却很有可能不能拒绝原假设。
- 另外，多重共线性是一个小样本问题，当 N 变大时，多重共线性导致的问题都可以被有效解决。
- 综合以上，多重共线性在小样本且系数不显著的情况下，可能是比较棘手的问题。

多重共线性的诊断

- 方差膨胀因子
- 另一个比较有效的方法时计算矩阵 $X'X$ 的条件数 (condition number)
 - 矩阵 $X'X$ 经过标准化之后，最大特征值与最小特征值之比即条件数。
 - 条件数越大，代表变量之间的相关性越严重，共线性也越严重
 - 一般认为条件数大于20时需要关注多重共线性的问题。

多重共线性的处理

- 如果存在多重共线性，通常非常难以处理：
 - 理论上最行之有效的方法是扩大样本量，然而在观测数据中通常是不可行的。
 - 常用的其他方法，比如删除变量、主成分分析等方法，都可能带来更加严重的问题。
 - 此外，使用Lasso或者岭回归方法会使得系数不可避免地产生偏差，因而如果我们关心的是因果推断，除非在高维情况下，这些方法也应该尽量避免使用。
- 幸运的是，只有小样本且核心解释变量不显著时，才需要担心多重共线性问题。
- 在实际应用中应该尽量使用观测更多或者更加微观的数据，结论会更加稳定。

受限最小二乘

- 有时回归模型中的系数可能存在天然的约束条件，此时我们可以通过约束最小二乘（constrained least squares）或者受限最小二乘（restricted least squares）的方法得到估计量。
- 为了简单起见，我们这里讨论带有线性约束的最小二乘估计问题，其中线性约束为：

$$R\beta - q = 0$$

其中 R 为 $r \times K$ 的（行满秩）矩阵， r 为约束个数， q 为 $r \times 1$ 的矩阵。

线性约束举例

多个约束

如果在回归：

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

如果理论上满足 $\beta_1 = 0, \beta_2 + \beta_3 = 1$ ，那么以上两个约束可以使用：

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}, q = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

带入 $R\beta + q = 0$ 来表示，此时 $r = 2$ ，即有两个约束。

受限最小二乘

- 我们将 $R\beta - q = 0$ 作为最小二乘法目标函数的约束条件，即可解得受限最小二乘估计量：

$$\begin{aligned} \min_{\beta} (Y - X\beta)'(Y - X\beta) \\ s.t. \quad R\beta - q = 0 \end{aligned}$$

- 通过拉格朗日乘子法，以上最优化问题等价于最小化：

$$\mathcal{L}(\beta, \lambda) = (Y - X\beta)'(Y - X\beta) - 2\lambda'(R\beta - q)$$

其中 2λ 为 $r \times 1$ 的拉格朗日乘子，因为有 r 个约束，相应的有 r 个拉格朗日乘子。

- 其一阶条件为：

$$\begin{cases} 2X'X\beta - 2X'Y - 2R'\lambda = 0 \\ 2(R\beta - q) = 0 \end{cases}$$

受限最小二乘

- 将第一个式子带入到第二个式子中，得到：

$$R \left[(X'X)^{-1} (X'Y + R'\lambda) \right] - q = 0$$

即：

$$\lambda = \left[R (X'X)^{-1} R' \right]^{-1} \left(q - R\hat{\beta} \right)$$

其中 $\hat{\beta}$ 为不受限的普通最小二乘估计量。

- 将上式带回到第一个式子中，就可以解得：

$$\hat{\beta}_c = \hat{\beta} - (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} (R\hat{\beta} - q)$$

以上便是在约束 $R\beta - q = 0$ 的条件下得到的最小二乘估计量。

联合检验

- 以上我们介绍了单个系数的假设检验问题，然而在许多应用中，我们需要对多个系数进行联合检验
- 对回归方程：

$$y_i = x_i' \beta + u_i$$

我们希望检验在真实的数据生成过程中： $R\beta - q = 0$ 是否成立，即检验原假设：

$$H_0 : R\beta - q = 0$$

以及备择假设 $R\beta - q \neq 0$ 。

联合检验的构造

- 为了构造检验统计量，我们可以分别讨论原假设成立以及备择假设成立条件下不受限的最小二乘估计量 $\hat{\beta}$ 以及首先最小二乘估计量 $\hat{\beta}_c$ 之间的差异。
- 如果原假设成立，即 $R\beta - q = 0$ 成立，如果记 $\hat{u}$ 为无约束下最小二乘估计量 $\hat{\beta}$ 计算的残差， $\hat{u}_c$ 为受限最小二乘估计量计算得到的残差，那么如果原假设成立，应当有 $\hat{u}_c' \hat{u}_c - \hat{u}' \hat{u} \approx 0$ 。
- 然而当原假设不成立时， $\hat{\beta}_c \neq \hat{\beta}$ ，此时两个残差平方和应该有非常大的差异，或者 $\hat{u}_c' \hat{u}_c \gg \hat{u}' \hat{u}$ 。

F统计量

- 为了进行假设检验，我们必须推导出在原假设成立的条件下， $\hat{u}'_c \hat{u}_c - \hat{u}' \hat{u}$ 的分布情况。
- 注意到以上的残差之差一定依赖于误差项的方差，所以为了构造一个基准 (pivotal) 统计量，我们还可以继续将以上的残差平方和除以误差项方差的估计，即 $s^2 = \hat{u}' \hat{u} / (N - K)$ 。
- 为此，如果假设 $u|x \sim \mathcal{N}(0, \sigma^2 I)$ ，假设原假设 $R\beta - q = 0$ 成立，我们有如下结论：

$$\frac{(\hat{u}'_c \hat{u}_c - \hat{u}' \hat{u}) / r}{\hat{u}' \hat{u} / (N - K)} \sim F(r, N - K)$$

(证明见讲义)

F检验

- 而如果备择假设成立，显然 $\hat{u}'_c \hat{u}_c - \hat{u}' \hat{u}$ 应该显著大于0。
- 为此，我们可以使用以上定理结论，通过计算： $F = \frac{(\hat{u}'_c \hat{u}_c - \hat{u}' \hat{u})/r}{\hat{u}' \hat{u}/(N-K)}$ 使用右侧检验：
 - 给定显著性水平 α ，当 $F > F_{\alpha}^{(r, N-K)}$ 时，即拒绝原假设，
 - 其中 $F_{\alpha}^{(r, N-K)}$ 为 $F(r, N-K)$ 分布的 $(1-\alpha)$ 分位数，即右侧 α 的临界值。

F统计量

- 以上介绍的 F 统计量在计算时有一个不方便的地方在于，为了计算 F 统计量，需要计算受约束和不受约束的回归，即进行两次回归。
- 然而实际上，只计算不受约束的回归也可以得到 F 统计量，如此可以避免计算受限最小二乘，从而减少计算量
- 以上检验统计量中，分母 $\hat{u}'\hat{u}/(N-K) = s^2$ ，而分子上，由于

$$\hat{u}_c = \hat{u} + X (X'X)^{-1} R' \left[R (X'X)^{-1} R' \right]^{-1} (R\hat{\beta} - q)$$

F统计量

- 从而

$$\begin{aligned}\hat{u}'_c \hat{u}_c &= \hat{u}' \hat{u} + \left(R\hat{\beta} - q \right)' \left[R \left(X'X \right)^{-1} R' \right]^{-1} \left(R\hat{\beta} - q \right) \\ &+ \hat{u}' X \left(X'X \right)^{-1} R' \left[R \left(X'X \right)^{-1} R' \right]^{-1} \left(R\hat{\beta} - q \right) \\ &+ \left(R\hat{\beta} - q \right)' \left[R \left(X'X \right)^{-1} R' \right]^{-1} R \left(X'X \right)^{-1} X' \hat{u} \\ &= \hat{u}' \hat{u} + \left(R\hat{\beta} - q \right)' \left[R \left(X'X \right)^{-1} R' \right]^{-1} \left(R\hat{\beta} - q \right)\end{aligned}$$

F统计量

- 从而

$$\hat{u}'_c \hat{u}_c - \hat{u}' \hat{u} = (R\hat{\beta} - q)' [R(X'X)^{-1}R']^{-1} (R\hat{\beta} - q)$$

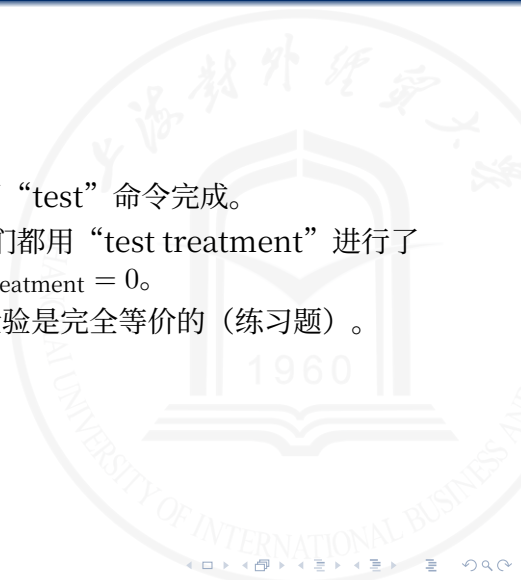
因而F统计量也可以写为

$$\begin{aligned} F &= \frac{(R\hat{\beta} - q)' [R(X'X)^{-1}R']^{-1} (R\hat{\beta} - q) / r}{s^2} \\ &= \frac{(R\hat{\beta} - q)' [R \{s^2 (X'X)^{-1}\} R']^{-1} (R\hat{\beta} - q)}{r} \end{aligned}$$

- 注意以上公式中，并不需要计算 $\hat{\beta}_c$

Stata中的 F 检验

- F 检验在Stata中可以在回归结束后直接使用“test”命令完成。
- 比如在OHIE的例子中，每次回归阶数后我们都用“test treatment”进行了对单个参数的 F 检验，其原假设为： $H_0 : \beta_{\text{treatment}} = 0$ 。
- 实际上，使用以上的 F 检验和单个系数的 t 检验是完全等价的（练习题）。



拟合优度检验

- 在线性回归的预测和拟合中，可以使用可决系数 R^2 度量拟合优度
- 更进一步的，我们有时还希望对模型的拟合优度进行检验。
- 最常见的检验是，当使用 x 对 y 进行预测时，其预测效果是否比仅仅使用平均值 $\bar{y}$ 对 y 进行预测究竟更好，或者 x 是否为对 y 的预测带来了更多的信息。
- 在线性回归的情景下，该检验相当于检验是否所有的 β_j （除常数项）都同时等于0。如果设截距项为 β_1 ，那么我们需要检验的原假设为：

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

方差分析

- 方差分析 (Analysis of Variance, ANOVA) 是在实验数据中常用的统计方法之一。
- 在实验中，通常将实验个体分为很多个组进行实验，并观察每个组的结果变量 y 是否有显著的差异。
- 在该方法中，我们将实验中的分类变量称之为因子 (factor)，而分类变量的取值称之为水平 (level)
 - 比如，在粮食优惠券的实验 (Jensen和Miller, 2011) 中，作者将低收入家庭分为了四组：不给优惠券；给予0.1元、0.2元、0.3元每斤的优惠券
 - 在这里，因子即家庭持有何种优惠券，而水平分为0元、0.1元、0.2元、0.3元每斤四种
 - 方差分析的目的即研究这四组的结果变量（比如营养摄入）是否有显著的不同。

方差分析与线性回归

- 一般地, 对于一个分组变量 $G_i \in \{1, 2, \dots, K\}$, 即 K 个分组, 令 $d_{ik} = \mathbb{1}\{G_i = k\}$ 为虚拟变量, 可以使用回归:

$$y_{ik} = \beta_1 + \sum_{k=2}^K \beta_k d_{ik} + u_{ik}$$

对 β_k 进行估计, 并使用 F 检验对原假设 $H_0: \beta_k = 0, k = 2, \dots, K$, 即每个组别的均值都相等, 进行检验。

- 根据上一节的结论, 以上检验的检验统计量可以写为:

$$F = \frac{ESS/(K-1)}{RSS/(N-K)}$$

在原假设的条件下 $F \sim F(K-1, N-K)$ 。

方差分析

- 注意到，在上述最小二乘回归中，由于只有虚拟变量作为解释变量，因而对于 y_{ik} 的最优预测即其所在组别的均值：如果 $d_{ik} = 1$ ，那么 $\hat{y}_{ik} = \bar{y}_k$ ，因而：

$$\begin{cases} TSS = \sum_{i=1}^N (y_{ik} - \bar{y})^2 = Y' M_0 Y \\ RSS = \sum_{i=1}^N \sum_{k=1}^K \left[d_{ik} (y_{ik} - \bar{y}_k)^2 \right] \\ ESS = \sum_{i=1}^N \sum_{k=1}^K \left[d_{ik} (\bar{y}_k - \bar{y})^2 \right] = \sum_{k=1}^K \left[N_k (\bar{y}_k - \bar{y})^2 \right] \end{cases}$$

- 由于 RSS 为 y_{ik} 减去相应组别均值 $\bar{y}_k$ 的平方和，因而通常也被称为组内平方和（sum of squares within groups），而 ESS 度量了不同组别之间均值之间的差异，因而也被称为组间平方和（sum of squares between groups）。

方差分析举例

不同教育水平的收入

- 作为示例，我们使用CFPS数据检验不同教育水平的对数收入均值是否相等 (oneway_anova.do)
- 可以看到te4变量共有7个分组， $K = 7$ ，样本量为 $N = 3226$ 。使用最小二乘法的 F 检验中， $F = 4.42$ ，残差平方和为 $RSS = 72794.67$ ；使用方差分析得到的结果中， $F = 4.42$ ， $RSS = 72764.67$ ，结果完全相同。

方差分析：更多比较

当然，我们也可以使用 F 检验单独比较某两个、三个组之间的均值是否相等，或者其他任意线性原假设。比如：

不同教育水平的收入

- 使用单个参数的 t 检验，比如使用文盲/半文盲（te4=1）作为基准组，并检验小学学历（te4=2）前面的系数是否为0：

```
1 reg log_income ib1.te4
```

- 使用 F 检验直接检验小学学历（te4=2）前面的系数等于0：

```
1 reg log_income ib1.te4
2 test 2.te4=0
```

- 使用所有的虚拟变量，并删掉常数项的回归，并直接检验 $H_0 : \beta_1 = \beta_2$ ：

```
1 reg log_income ibn.te4, noconstant
2 test 1.te4=2.te4
```


结构变化检验

- 解决以上问题的办法是想办法将其转化为一个回归。实际上，以上回归等价于：

$$y_i = (d_i \cdot x_i)' \beta_1 + [(1 - d_i) \cdot x_i]' \beta_0 + u_i$$

即将 x 的每个分量（包括常数项）都乘以 d_i 和 $1 - d_i$ 进行回归。

- 因而，为了检验原假设： $H_0 : \beta_0 = \beta_1$ ，可以直接估计上式，并使用 F 检验即可。
- 注意由于 β_0, β_1 均为 K 维向量，因而在该原假设中，有 $2K$ 个参数以及 K 个约束，因而 F 分布的自由度应为 $F(K, N - 2K)$ 。以上检验由邹至庄于1960年提出（Chow, 1960），因而也被称为邹氏检验（Chow test）。

邹氏检验实例

比较不同城市回归

我们使用如下回归方程比较一二线城市（北上广深+省会城市）与其他城市的人口增长模式是否有不同：

$$y_i = \beta_{01} (1 - d_i) + \beta_{11}d + \beta_{02} (1 - d_i) \cdot x_{i2} + \beta_{12}d \cdot x_{i2} + \beta_{03} (1 - d_i) \cdot x_{i3} + \beta_{13}d_i \cdot x_{i3} + u_i$$

其中 y 为城市的人口自然增长率, x_2 为人口对数, x_3 为人口密度对数, 该比较的原假设为:

$$H_0 : \begin{cases} \beta_{01} = \beta_{11} \\ \beta_{02} = \beta_{12} \\ \beta_{03} = \beta_{13} \end{cases}$$

chow_test.do

邹氏检验

- 或者，以上检验式也可以写为：

$$y_i = x_i' \beta_0 + d_i \cdot x_i' (\beta_1 - \beta_0) + u \triangleq x_i' \beta_0 + d_i \cdot x_i' \delta + u_i$$

其中 $\delta = \beta_1 - \beta_0$

- 从而对原假设： $H_0 : \delta = 0$ 进行检验也可以达到同样的目的
- 比如，使用该思路，上例的检验可以通过如下回归方程完成：

$$y_i = \gamma_1 + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \delta_1 d_i + \delta_2 d_i \cdot x_{i2} + \delta_3 d_i \cdot x_{i3} + u_i$$

我们需要检验原假设： $H_0 : \delta_1 = \delta_2 = \delta_3 = 0$ ：

```
1 reg pop_growth log_pop log_pop_dens d d_log_pop d_log_pop_dens
2 test d d_log_pop d_log_pop_dens
```


大样本下的检验

- 以上我们讨论了多个系数的联合检验问题，然而，以上的检验方法严重依赖于同方差假设以及正态性假设，而这些是非常强的假设。
- 一个很自然的问题是，我们能不能像在单个系数的假设检验中那样，放弃正态分布假设以及同方差假设进行联合检验呢？答案是可以的。
- 在这种情况下，我们一般可以使用大样本的Wald检验。

线性原假设下的Wald检验

- 我们首先考虑线性原假设: $H_0 : R\beta - q = 0$
- 在同方差下, 有:

$$\sqrt{N} (\hat{\beta} - \beta) \stackrel{a}{\sim} \mathcal{N} \left(0, \sigma^2 [\mathbb{E} (x_i x_i')]^{-1} \right)$$

- 根据以上结论, 我们可以计算:

$$R\hat{\beta} - q \stackrel{a}{\sim} \mathcal{N} \left(R\beta - q, \frac{\sigma^2 R [\mathbb{E} (x_i x_i')]^{-1} R'}{N} \right)$$

从而在原假设 $R\beta - q = 0$ 的条件下:

$$R\hat{\beta} - q \stackrel{a}{\sim} \mathcal{N} \left(0, \frac{\sigma^2 R [\mathbb{E} (x_i x_i')]^{-1} R'}{N} \right)$$

线性原假设下的Wald检验

- 两边同时乘以 $\left[\frac{\sigma^2 R [\mathbb{E}(x_i x_i')]^{-1} R'}{N} \right]^{-1/2}$ 得到:

$$\left[\frac{\sigma^2 R [\mathbb{E}(x_i x_i')]^{-1} R'}{N} \right]^{-1/2} (R\hat{\beta} - q) \sim \mathcal{N}(0, I_{r \times r})$$

注意由于 R 为 $r \times K$ 维, 因而 $(R\hat{\beta} - q)$ 是一个 $r \times 1$ 的向量。

- 将其与其转置相乘, 就得到了

$$(R\hat{\beta} - q)' \left[\frac{\sigma^2 R [\mathbb{E}(x_i x_i')]^{-1} R'}{N} \right]^{-1} (R\hat{\beta} - q) \sim \chi^2(r)$$

线性原假设下的Wald检验

- 然而，注意到以上方程中， σ^2 和 $\mathbb{E}(x_i x_i')$ 是未知的，因而可以使用插入法，使用 s^2 替代 σ^2 ，使用 $1/N \sum_{i=1}^N (x_i x_i') = X'X/N$ 代替 $\mathbb{E}(x_i x_i')$ ，即可得到：

$$\left(R\hat{\beta} - q \right)' \left[R \left\{ s^2 (X'X)^{-1} \right\} R' \right]^{-1} \left(R\hat{\beta} - q \right) \stackrel{a}{\sim} \chi^2(r)$$

- 比较以上的检验统计量与 F 检验的检验统计量，有：

$$\left(R\hat{\beta} - q \right)' \left[R \left\{ s^2 (X'X)^{-1} \right\} R' \right]^{-1} \left(R\hat{\beta} - q \right) = rF$$

因而我们通常也将其写为 $rF \stackrel{a}{\sim} \chi^2(r)$ 。

线性原假设下的Wald检验

- 在原假设成立的条件下，检验统计量 $rF \approx 0$
- 然而当备择假设成立时，检验统计量 rF 应该显著大于0
- 因而再次，我们可以使用右侧检验对原假设进行检验。
- 实际上，由于 F 分布的极限分布（第2个自由度趋向于无穷时）为 χ^2 分布，因而在同方差的条件下，以上 χ^2 检验可以直接使用之前的 F 检验的结果，只不过现在 χ^2 检验并不需要误差项服从正态分布这一假设。

线性原假设下的Wald检验

- 而在异方差的条件下，推理与上面的方法类似，我们仅仅需要将同方差下 $\hat{\beta}$ 的协方差矩阵的估计： $s^2 (X'X)^{-1}$ 替换为异方差稳健的估计：

$$\widehat{\mathbb{V}}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{i=1}^N \hat{u}_i^2 x_i x_i' \right) (X'X)^{-1}$$

从而使用：

$$(R\hat{\beta} - q)' \left[R \left\{ \widehat{\mathbb{V}}(\hat{\beta}) \right\} R' \right]^{-1} (R\hat{\beta} - q) \stackrel{a}{\sim} \chi^2(r)$$

即可

- 在Stata中，使用reg命令时加入选项robust，所有的 F 检验都将使用异方差稳健的协方差矩阵估计。

非线性原假设

- 以上所有讨论都建立在线性原假设 $H_0 : R\beta - q = 0$ 的基础上，而更加一般的，有时我们会对一个非线性的原假设进行假设检验。
- 比如，我们可能会对原假设 $H_0 : \beta_1^2 - 1 = 0$ 做假设检验，该假设检验相当于检验了：

$$H_0 : \beta_1 = 1 \text{ or } \beta_1 = -1$$

- 尽管是两个等式检验，但是由于两个命题是相互排斥的，因而无法使用线性原假设的检验方法进行检验。
- 般地，我们记原假设为： $H_0 : R(\beta) - q = 0$ 其中 $R(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}^r$ 为一个向量函数：

$$R(\beta) = \begin{bmatrix} R_1(\beta) \\ \vdots \\ R_r(\beta) \end{bmatrix}$$

非线性原假设下的Wald检验

- 同样，我们使用Wald检验的思路，从 $\hat{\beta}$ 的渐进分布出发，记：

$$(\hat{\beta} - \beta) \stackrel{a}{\sim} N(0, \mathbb{V}(\hat{\beta}))$$

- 在同方差假设下，

$$\mathbb{V}(\hat{\beta}) = \frac{\sigma^2 [\mathbb{E}(x_i x_i')]^{-1}}{N}$$

- 在异方差假设下，

$$\mathbb{V}(\hat{\beta}) = \frac{[\mathbb{E}(x_i x_i')]^{-1} \mathbb{E}(u_i^2 x_i x_i') [\mathbb{E}(x_i x_i')]^{-1}}{N}$$

非线性原假设下的Wald检验

- 如果原假设成立，那么 $R(\hat{\beta}) - q \approx 0$ ，我们不妨对使用Delta方法，将上式在 β 处进行泰勒展开：

$$R(\hat{\beta}) - q = R(\beta) - q + \frac{\partial R(\beta)}{\partial \beta'} (\hat{\beta} - \beta) + O\left(|\hat{\beta} - \beta|^2\right)$$

记

$$\dot{R} = \frac{\partial R(\beta)}{\partial \beta'} = \begin{bmatrix} \frac{\partial R_1(\beta)}{\partial \beta'} \\ \vdots \\ \frac{\partial R_r(\beta)}{\partial \beta'} \end{bmatrix}_{r \times K}$$

- 在原假设成立的条件下，有：

$$[R(\hat{\beta}) - q] = \dot{R}(\hat{\beta} - \beta) + o_p(1) \xrightarrow{D} \dot{R}(\hat{\beta} - \beta) \stackrel{d}{\sim} N(0, \dot{R} \mathbb{V}(\hat{\beta}) \dot{R}')$$

非线性原假设下的Wald检验

- 同样，如果 $\text{rank}(\dot{R}V\dot{R}') = r$ ，我们可以得到：

$$(\dot{R}V(\hat{\beta})\dot{R}')^{-\frac{1}{2}} [R(\hat{\beta}) - q] \stackrel{a}{\sim} \mathcal{N}(0, I_r)$$

或者：

$$[R(\hat{\beta}) - q]' (\dot{R}V(\hat{\beta})\dot{R}')^{-1} [R(\hat{\beta}) - q] \stackrel{a}{\sim} \chi^2(r)$$

其中：

- $V(\hat{\beta})$ 可以使用之前介绍的协方差矩阵估计
- $\dot{R}$ 未知，但是由于 $R(\cdot)$ 的函数形式知道，因而我们可以同样使用插入法： $\hat{\dot{R}} = \frac{\partial R(\hat{\beta})}{\partial \beta'}$ 进行估计
- 从而，最终的检验统计量为：

$$[R(\hat{\beta}) - q]' (\hat{\dot{R}} \{ \widehat{V(\hat{\beta})} \} \hat{\dot{R}}')^{-1} [R(\hat{\beta}) - q] \stackrel{a}{\sim} \chi^2(r)$$

以上就是任意原假设条件下的Wald检验统计量。

非线性Wald检验示例

非线性原假设的检验

在OLS中，如果原假设为 $H_0 : \beta_k^2 - 1 = 0$ ，那么根据Wald检验统计量构造的方法，首先根据大样本性质：

$$\left(\hat{\beta}_k - \beta_k \right) \overset{a}{\sim} \mathcal{N} \left(0, \mathbb{V} \left(\hat{\beta} \right)_{kk} \right)$$

将 $\hat{\beta}_k$ 带入到原假设中，并在 β_k 处进行泰勒展开：

$$\begin{aligned} \left(\hat{\beta}_k^2 - 1 \right) &= \left(\beta_k^2 - 1 \right) + 2\beta_k \left(\hat{\beta}_k - \beta_k \right) + O \left(\left(\hat{\beta}_k - \beta_k \right)^2 \right) \\ &= 2\beta_k \left(\hat{\beta}_k - \beta_k \right) + O \left(\left(\hat{\beta}_k - \beta_k \right)^2 \right) \\ &= 2\beta_k \left(\hat{\beta}_k - \beta_k \right) + o_p(1) \\ &\overset{a}{\sim} \mathcal{N} \left(0, 4\beta_k^2 \mathbb{V} \left(\hat{\beta} \right)_{kk} \right) \end{aligned}$$

非线性Wald检验示例

非线性原假设的检验

从而：

$$\frac{(\hat{\beta}_k^2 - 1)}{2|\beta_k| \sqrt{\mathbb{V}(\hat{\beta})_{kk}}} \stackrel{a}{\sim} \mathcal{N}(0, 1)$$

或者：

$$\frac{(\hat{\beta}_k^2 - 1)^2}{4\beta_k^2 \mathbb{V}(\hat{\beta})_{kk}} \stackrel{a}{\sim} \chi^2(1)$$

使用 $\hat{\beta}_k$ 代替 β_k ，使用 $\widehat{\mathbb{V}(\hat{\beta})_{kk}}$ 代替 $\mathbb{V}(\hat{\beta})_{kk}$ ，得到：

$$\frac{(\hat{\beta}_k^2 - 1)^2}{4\hat{\beta}_k^2 \widehat{\mathbb{V}(\hat{\beta})_{kk}}} \stackrel{a}{\sim} \chi^2(1)$$

以上即我们需要的检验统计量。

