回归分析在计量经济学中的应用

3180104294　孔畅　2020年1月12日

前言

回归分析在很多领域都有着大量的应用，尤其是在经济学领域，以回归分析和时间序列分析为基础诞生了计量经济学。计量的意思就是以统计方法做定量研究，是一整套建立在统计学上的分析研究。

经济变量之间往往存在着相关性，甚至有着统计依赖关系，对于统计依赖关系的考察主要就是通过回归分析来完成的。本文将通过计量经济学来展示回归分析在具体学科中的应用，利用具体应用换一个角度重新看待回归分析。

普通线性回归模型及检验

对于经济学而言，这个世界上的经济变量之间，最为简单也是最为常见的关系就是线性相关了，因此研究线性回归模型是计量经济学的基础。

普通最小二乘回归

假设变量之间存在线性关系，再增加一个随机扰动项，得到总体回归函数（PRF）：

$y = X\beta + e$

通过观测样本数据，寻找样本回归函数（SRF）来进行拟合，得到β_0和β_1的估计值。最常用的方法就是普通最小二乘法（OLS）。

再假设误差项e_i满足零均值同方差不相关三个条件，即可通过最小二乘法得到β的无偏有效点估计。

假设：

$E(e_i) = 0$

$Var(e_i) = \sigma^2$

$Cov(e_i, e_j) = 0$

可得：

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\hat{\sigma}^2 = \frac{RSS}{n-p}$$

同时可以验证β的最小二乘估计与极大似然估计结果相同，但需要注意σ的最小二乘估计则小于极大似然估计。

其中为了方便，分别定义总体平方和（TSS），回归平方和（ESS）以及残差平方和（RSS）：

$$TSS = \sum y_i^2$$

$$ESS = \sum \hat{y}_i^2$$

$$RSS = \sum e_i^2 = y'y - \hat{\beta}' X'y = TSS - ESS$$

从而我们可以得到一种衡量拟合程度的统计量，可决系数R^2：

$$R^2 = ESS/TSS$$

可决系数说明回归结果在多大程度上是可以解释的，能大致上判断回归的拟合程度，不过在复杂的现实中存在一些缺陷。对于计量经济学来说，大部分的数据首先要做的就是普通线性回归，在应用中我们还要对回归结果进行显著性检验。

假设检验

大部分的参数显著性检验都可以变换为两种检验：t检验和F检验。t检验用来检验单个参数的显著性，而F检验用来检验整个方程的显著性，二者都需要用到约束最小二乘估计。

约束最小二乘估计中，在满足条件$A\beta = b$的情况下，β的最小二乘估计为：

$$\hat{\beta}_c = \hat{\beta} - (X'X)^{-1}A(A(X'X)^{-1}A')^{-1}(A\hat{\beta} - b)$$

但在进行显著性检验时，可以将约束条件代入从而约简模型，变成无约束普通最小二乘估计：

$\hat{\alpha} = (Z'Z)^{-1}Z'y$，　　其中$Z$是受到约束的$X$

从而得到新的$RSS_H = y'y - \hat{\alpha}Z'y$，通过构造以下统计量进行检验：

$$F_H = \frac{(RSS_H - RSS)/m}{RSS/(n-p-1)} \sim F(m, n-p-1)$$

由于F检验的假设为$H:\beta = 0$，所以代入得到：

$$F_H = \frac{ESS/p}{RSS/(n-p-1)} \sim F(p, n-p-1)$$

再根据显著性水平查表选择是否接受原假设即可。

F检验能说明整个回归方程是否显著，但我们不知道其中单个变量是否对于结果显著，这时候要选择使用t检验。假设$H:\beta_i = 0$，通过代入计算得到：

$$F_H = \frac{\hat{\beta}_i^2}{\hat{\sigma}^2 c_{i+1,i+1}} \sim F(1, n-p-1)$$，　其中$c_{i+1,i+1}$是$(X'X)^{-1}$的第$i+1$个对角元

再通过t分布与F分布的关系构建t统计量：

$$t_i = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{c_{i+1,i+1}}} \sim t(n-p-1)$$

最后根据显著性水平查表选择是否接受原假设即可。其中$\hat{\sigma}\sqrt{c_{i+1,i+1}}$也被称为$\hat{\beta}_i$的标准误差，各种计量经济学的分析软件都会自动计算该值。

至此，回归分析在计量经济学中最基本的应用就完成了，包括了从模型建立到假设检验的完整过程，但现实中的经济数据通常不能很好地满足最开始的假设条件，引发了各种各样的问题需要解决。

多重共线性

对于截面数据，最常见的问题就在于多重共线性，即$Cov(x_i, x_j) \neq 0$，不同的解释变量之间存在相关性。例如研究一个人身高体重对于体育成绩的相关性，身高与体重之间本身就具有相关性。

后果

如果x_i之间存在线性关系，则称为存在完全共线性（perfect multicollinearity），这种情况下$(X'X)^{-1}$根本不存在，完全无法得到参数的估计量。而如果并不是完全的共线性，也会导致$(X'X)^{-1}$的主对角线元素较大，引起方差的增大，参数估计量非有效：

$$Var(\hat{\beta}_i) = \frac{\sigma^2}{\sum x_{1i}^2} * \frac{1}{1-r^2} > \frac{\sigma^2}{\sum x_{1i}^2}$$

我们称$1/(1-r^2)$为方差膨胀因子（VIF）。在$r \neq 0$的情况下，参数估计的方差将会增大，使得参数在进行t检验时更容易接受假设$H: \beta_i = 0$，从而将某些重要变量排除在模型之外，最终导致模型失效。

检验

首先，我们要想办法判断一个模型是否存在多重共线性。

对于只有两个解释变量的模型，只需要简单计算相关系数即可。而对于多个解释变量的模型，我们需要用到综合统计方法，如果某个模型的F检验值较大，而t检验值较小，说明模型很可能存在多重共线性。

在判断出某个模型存在多重共线性后，接下来我们要确认是哪些变量引起了多重共线性。最常用的方法就是逐步回归法，一个一个将解释变量引入模型中，每次观察可决系数的变化。但随着引入的变量增多，可决系数必然增大，于是我们需要一种新的可决系数来判断新加入的变量是否对于模型的精度有显著提高。

通过对原先的可决系数进行自由度加权调整，我们得到新的调整可决系数：

$$\overline{R}^2 = 1 - \frac{RSS/(n-q-1)}{TSS/(n-1)} = 1 - \frac{n-1}{n-q-1}(1-R^2)$$

当加入一个新变量后，如果调整可决系数下降，说明新加入变量对于模型精度提高不显著，很可能对模型没有影响或者与某些已经引入的解释变量高度相关。

另一种方法是进行辅助回归（Auxiliary Regression），通过对某个变量x_i用$x_j (j \neq i)$进行线性回归，如果可决系数较大说明其与其他变量相关。

修正

找出具有多重共线性的变量后，下一步需要想办法克服多重共线性，最常用的方法是直接排除引起多重共线性的变量，以逐步回归法得到最广泛的应用。

也可以放弃普通最小二乘估计，改用其他估计方法来减小方差，例如岭回归。因为多重共线性的主要后果在于回归结果的方差增大，所以通过人为引入偏误的方法可以减小参数估计量的方差：

$\hat{\beta} = (X'X + D)^{-1}X'y$，　　其中$D$通常选择主对角阵

显然新的估计量具有更小的方差。

当然在现代计量经济学的很多研究中，多重共线性的解决方案十分简单粗暴：增大样本数量。这也是最有效的方法。

异方差性

除了多重共线性，还有一种常见的问题是异方差性，即$Var(e_i) = f(x) * \sigma^2$。例如研究家庭储蓄行为，高收入家庭的储蓄方差显然要比低收入家庭更大。

后果

由于最小二乘估计中利用了$E(ee') = \sigma^2 I$，所以OLS结果将不具有有效性，但不影响无偏性。由于方差的变化，t检验也将随之失去意义，其他检验也将受到类似影响。同时模型的置信区间将增大，降低预测精度。

检验

最原始的检验异方差性的方法就是直接进行OLS估计，随后画图判断方差是否有某种非常数函数的趋势。

现代较为常用的方法是布罗施-帕甘（Breusch-Pagan）检验。如果某个模型不存在异方差性，那么我们构建辅助回归模型：

$e_i^2 = X\delta + \epsilon_i$，　　其中$\epsilon_i$是同方差随机误差项

假设为$H: \delta = 0$。通过对辅助回归模型进行F检验或者拉格朗日乘数检验（LM检验），可以判断误差项e_i是否与解释变量高度相关。

计量经济学在后来还发展出怀特（White）检验，是布罗施-帕甘检验的一种拓展。其基本思路与布罗施-帕甘检验类似，先进行OLS回归，得到误差项e_i^2的估计值，之后构建辅助回归（以二元为例）：

$$e_i^2 = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i1}^2 + \alpha_4 x_{i2}^2 + \alpha_5 x_{i1} x_{i2} + \epsilon_i$$

通过辅助回归的可决系数$nR^2 \sim \chi^2(h)$检验结果。

修正

在确定异方差性后，通常使用加权最小二乘法（WLS）来解决问题。用$\sqrt{f(x)}$去除原模型，即得：

$$\frac{y}{\sqrt{f}} = \frac{X\beta}{\sqrt{f}} + \frac{e}{\sqrt{f}}$$

由此新的$Var(e^*) = \sigma^2$，消除了异方差性。

但我们并不知道e的具体值，因此要先对其进行估计。最简单的方法是直接使用OLS估计出：

$\sigma^2 \hat{W} = diag\{\hat{e}_1^2, \hat{e}_2^2, \dots, \hat{e}_n^2\}$

由此直接使用$D^{-1} = diag\{1/|\hat{e}_1|, 1/|\hat{e}_2|, \dots, 1/|\hat{e}_n|\}$作为权矩阵即可，但采用这种方法有时会显得粗糙。

计量经济学中广泛应用的方法是，假设方差有着如下指数形式：

$e_i^2 = \sigma^2 e^{X\alpha}$

$ln(e_i^2) = X\delta + v_i,\quad$ 其中$\delta_0 = ln\sigma^2 + \alpha_0$

对新的模型进行估计得：

$\hat{\sigma}_i^2 = e^{X\hat{\alpha}}$

$\hat{w}_i = 1/\sqrt{e^{X\hat{\alpha}}}$

但进行加权最小二乘法时，找到正确的函数关系绝非易事，因此更为一般的方法是直接修正方差估计值，因为异方差性仅仅影响估计的有效性，而不影响无偏性和一致性。例如在一元模型中：

$$Var(\hat{\beta}_1) = \frac{\sum x_i^2\sigma_i^2}{(\sum x_i^2)^2}$$

使用e_i替代后，新得到的标准差即为异方差稳健标准误，这种方法也被称为异方差稳健标准误法，是在计量经济学应用中解决异方差性的主要手法。

内生解释变量

内生解释变量问题通常分为三种情况：被解释变量与解释变量具有联立因果关系；模型设定时遗漏了重要变量；解释变量存在测量误差。

例如研究企业引进外资是否提高了企业的实际效益，我们选择用企业外资比例和其他外生变量作为解释变量，建立对于被解释变量利润率的模型。但实际上，只有利润率高的企业才有比较好的机会来引进外资，因此存在联立因果关系。

又例如劳动者工资主要由受教育程度，工作经验，个人能力等诸多因素决定。但劳动者个人能力难以测定，因此进入随机误差项。然而个人能力和受教育程度密切相关，因此模型中会出现随机误差项与受教育程度高度相关的情况，遗漏了重要变量。

后果

出现内生解释变量后，不同性质的内生解释变量将产生不同的后果，通常会造成估计有偏。例如在一元线性回归中，如果解释变量与随机误差项存在正相关，β_0的估计值就会减小，β_1则会增大，负相关则相反。

修正

首先来考虑如何修正内生性。

最常用的方法是工具变量法。通过选取与随机误差项不相关，而与内生解释变量高度相关的工具变量（IV）来替代原来的内生解释变量，从而消除内生性。同时还应注意尽量不与其他解释变量相关，防止多重共线性。

拿一元回归模型举例，利用矩估计，在$E(e_i) = 0, E(x_i e_i) = 0$的条件下，得到正规方程组：

$\sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$

$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$

利用总体矩条件可得：

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}, \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

之后在$E(x_i e_i) = 0$不满足时引入工具变量z，将上述估计过程改为

$$\beta_1^* = \frac{\sum z_i y_i}{\sum z_i x_i}$$

这种方法即为工具变量法，相应变量称为工具变量法估计量。对于多元情况同样使用工具变量矩阵即可。

由于$\lim \frac{1}{n} \sum z_i e_i = Cov(z_i, e_i) = 0$，所以在大样本下$\beta_1^*$依概率收敛于$\beta_1$，工具变量法估计量具有渐进无偏性。

实际应用中通常先通过多个工具变量对替代变量进行OLS估计，然后再代入最初的模型中进行OLS估计，因此也被称为两阶段最小二乘法（2SLS）。

检验

通常来说，可以依靠经济学知识来判断哪些变量可能具有内生性。

另外，豪斯曼（Hausman）从数学角度给出了检验方法。中心思想为，依靠工具变量的估计结果与原估计结果比较，观察差异是否显著，显著则被替代变量具有内生性。

首先用工具变量对替代变量进行OLS估计得到残差：
$x_i = Z\alpha + v_i$

接着将残差项加入原模型中进行OLS估计：

$y = X\beta + \delta \hat{v}_i + e_i$

接着检验假设$H: \delta = 0$即可。

但这种检验方法引出了另一个问题，如何确保所有工具变量都是外生变量，这就要依靠过度识别检验（ORT）。

过度识别检验的基本思路是，通过直接进行2SLS回归，将残差项再关于工具变量和原模型中的外生变量进行OLS回归，新的回归中假设所有工具变量对于残差项无影响，利用F检验判断即可。

总结

通过对回归分析在计量经济学中的具体应用的学习，可以感受到理论与实际应用之间有着不小的差距，完美的假设在现实中几乎不存在，需要通过各种方法来修正出现的问题。仅仅是截面数据的回归分析就能存在着无数的问题，如果加入时间序列数据将会产生更多的问题。

这些问题需要我们借助相关领域的经验，灵活运用学到的知识，并不能简单地代入公式可得，需要具体情况具体分析。